A NOVEL METHOD FOR CLUSTER ANALYSIS OF RNA STRUCTURAL DATA

A Thesis
Presented to
The Academic Faculty

by

Emily Rogers

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Computational Science and Engineering

Georgia Institute of Technology
August 2018

A NOVEL METHOD FOR CLUSTER ANALYSIS OF RNA STRUCTURAL DATA

Approved by:

Dr. David Bader, Committee Chair
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Christine Heitsch, Advisor
School of Mathematics
Georgia Institute of Technology

Dr. Srinivas Aluru
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Roger Wartell
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Roman Aranda
Defense Forensic Science Center
Department of Defense

Date Approved: April 18, 2018

To my family

ACKNOWLEDGEMENTS

Thanks to my family, who supported in every way: emotionally, practically, financially, etc.

Thanks to my legion of babysitters, who often came to help at the drop of a hat. You know who you are.

Thanks to my committee, for the willingness in being an integral part of my degree.

And finally, a special thanks to my advisor Dr. Christine Heitsch. You took a chance on me, bore with me through the rough patches, tolerated the sleepless weeks when little work was being done, pushed me even when there was tears, taught me how to communicate precisely, gave me a pattern of success, and never ever gave up on me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This dissertation concerns application of data mining techniques to Boltzmann samples of RNA sequences, in addition to forensic DNA mixture interpretations. With the former, important structural information is mined from a set of RNA secondary structures, in order to not only predict structure but also to enable comparisons between samples for key similarities and differences. With the latter, through visualization and statistical analysis, we illuminate the state of forensic DNA mixture interpretation and show that high quality mixture interpretations are possible.

Chapter one contains the initial paper establishing RNA profiling with its proof of principle, published in *Nucleic Acids Research* in 2014. This paper lays out methodology of RNA profiling: the definition and calculation of an equivalence class on RNA helices, the selection of the top helix classes to form a set of features, the profiling of each secondary structure according to its set of features, and the selection and relation of the most frequent profiles into a summary profile graph. It also uses the motivating example of the small RNA `VcQrr3` to show the benefits of profiling: its ability to highlight the multimodal structural signal in a Boltzmann sample, which for `VcQrr3` correlates strongly with experimental data.

Chapter two contains the review paper comparing Boltzmann sampling methods (including RNA profiling) against the more popular minimum free energy (MFE) method, published in *Wiley Interdisciplinary Reviews: RNA* in 2016. This paper further establishes the benefit of sampling methods, demonstrating its prediction accuracy at the base pair level to be on average at least as good as that of MFE methods. Furthermore, it establishes the principle that higher levels of structural abstraction is correlated with higher prediction accuracies. Finally, it presents aptamer screening as the motivating example for the iterative use of abstraction methods to progressively refine prediction.

Chapter three presents additional work on RNA profiling, in an attempt to extend its

ability to extract meaningful structural signal from sequences longer than 300 nucleotides. It is motivated by the observation that many 'competing' helices in a Boltzmann sample are in fact occupying the same structural niche. Thus, a level of abstraction above profiling should be defined that combines all such competing-yet-similar helices into a higher level class. This enables the structural signal to emerge more clearly, i.e. the ability to see the major structural characteristics of the sample without getting bogged down in low level details. One such way to combine similar helices is based on calculating their conditional probabilities, and forming boolean logicals (AND/OR relations) between helices.

Chapter four comprises of the first application of profiling, specifically to the conditioning of RNA thermodynamic optimization under the thermodynamic model perturbation (published in *Biophysical Journal* in 2017). Profiling is used both to quantify a condition number, and to define the robustness of the Boltzmann sample to thermodynamic perturbation. We show a correlation between the mathematical condition number and the biologically defined notion of robustness, thus providing conditioning with intuitive thresholds for well- vs. ill-conditioning.

Chapter five comprises the second application of profiling, this time to the consensus structure problem. Given a set of homologous sequences, our novel method `ConsensusStems` uses profiling to find the common consensus structure that the sequences all fold to. By clustering the individual features from each sequence, the native structure signal is extracted with a degree of high accuracy. Its high accuracy and efficiency is due to its use of abstraction to turn messy low level structural information into a clear, strong consensus signal.

CHAPTER I

PROFILING SMALL RNA REVEALS MULTIMODAL SUBSTRUCTURAL SIGNALS IN A BOLTZMANN ENSEMBLE

This chapter is published in *Nucleic Acids Research* 42, no. 22 (2014): e171

1.1 Abstract

As the biomedical impact of small RNAs grows, so does the need to understand competing structural alternatives for regions of functional interest. Suboptimal structure analysis provides significantly more RNA base pairing information than a single minimum free energy prediction. Yet computational enhancements like Boltzmann sampling have not been fully adopted by experimentalists since identifying meaningful patterns in this data can be challenging. Profiling is a novel approach to mining RNA suboptimal structure data which makes the power of ensemble-based analysis accessible in a stable and reliable way. Balancing abstraction and specificity, profiling identifies significant combinations of base pairs which dominate low-energy RNA secondary structures. By design, critical similarities and differences are highlighted, yielding crucial information for molecular biologists. The code is freely available via `http://gtfold.sourceforge.net/profiling.html`.

1.2 Introduction

RNA molecules perform a variety of important functions, including the expanding roles of "small" RNAs [24, 36]. Short, non-coding RNA molecules are now known to function in chemical catalysis as ribozymes [11, 139], in aptamer binding as riboswitches [139, 160], and in the quorum sensing mechanism of bacteria like *Vibrio cholerae* [100, 158].

Knowing the base pairings of an RNA sequence is critical to understanding its function. A first step is often to predict a minimum free energy (MFE) secondary structure under the nearest neighbor thermodynamic model (NNTM). However, even for short sequences, the MFE prediction may not be the native secondary structure [106, 35].

Prediction accuracy improves when suboptimal structures are considered [169, 75, 182, 74, 181, 180]. Although they can be generated exhaustively [171] or sampled deterministically [178], the current standard is to sample structures stochastically from the Boltzmann distribution [32, 105]. The goal is to identify the set of base pairs which dominate the low-energy secondary structures and hence are more likely to occur in nature. The challenge is to extract the most meaningful structural signal from a noisy Boltzmann sample.

At the level of individual base pairs, this has been well-studied [109, 65, 71, 104]. It is known that, even when disjoint, two Boltzmann samples (typically of size 1000) will display "nearly identical patterns" of estimated probabilities [32]. Given the significance of high frequency pairings, it is natural to ask which *combinations* dominate the low-energy secondary structures.

High probability helices, with few low-energy competitors, are a structural signal strong enough to be identified by visual inspection of a 2D dot plot. However, beyond these well-determined regions, the signal is much less clear. In particular, there will be regions where one can easily see that competing structural alternatives exist, but not what they might be.

Clarifying this multimodal signal is critical to advancing our understanding of RNA structure and function. This is especially true for RNAs whose functionality may depend on switching from one conformation to another [139, 160]. However, identifying combinations of base pairs whose probability is high enough to merit attention but which have significant competing alternatives is challenging.

Existing methods [30, 49] identify dominant combinations of base pairs by dividing the Boltzmann sample into groups, and reporting a representative structure for each one. However, as illustrated below, support for different substructures can be lost within a group or diluted across groups. This poses obstacles to understanding the substructural signal in a Boltzmann ensemble, especially when multimodal.

Communicating significant commonalities and differences in pairing combinations is critical to understanding competing structural alternatives for regions of functional interest. Given this, we introduce a new combinatorial approach to analyzing a Boltzmann sample.

Our method focuses on denoising the distribution of helices; those with high enough probability form our set of "features" which are used to "profile" the structures. In this way, we identify notable combinations of helices and present this signal as concisely and stably as possible. By design, RNA profiling highlights critical relations at the substructure level, yielding crucial information for molecular biologists.

1.2.1 VcQrr3: a case study

As concrete motivation, we consider a small RNA sequence with an unknown structure from the pathogen *Vibrio cholerae*. This bacteria regulates its virulence via a quorum sensing mechanism [111, 175] that involves four short, non-coding RNA molecules, denoted VcQrr1–VcQrr4 [100]. With cholera infecting three million people and causing 100,000 deaths annually [115], understanding the structure and function of these small RNAs is an important biomedical problem [73].

Quorum regulatory RNA (Qrr) molecules have been found in multiple *Vibrio* species [100, 158, 112], and sequence alignment identifies a 32 nucleotide region which is essentially perfectly conserved. This degree of sequence conservation is strong evidence for functional significance; however it provides no structural information for the region of interest.

Moreover, thermodynamic optimization [185, 162] predicts that the four VcQrr sequences have three different MFE structures [100] with varying roles for the conserved region. Given this lack of structural consensus, it is important to consider a more nuanced view of base pairing alternatives.

Figure 1 shows the VcQrr3 MFE structure. As seen in Figure 2, base pair probabilities clearly support the formation of the first and fourth helices. However, the situation for the middle two, and most of the conserved region, is considerably murkier; we see that significant structural alternatives exist but not what they might be.

Parsing this multimodal structural signal requires analyzing the suboptimal structures from a Boltzmann sample. Understanding its nature requires preserving the critical relations. To appreciate the challenge, consider the suboptimal secondary structures for VcQrr3, denoted s_1, s_2, and s_3, shown in Figure 3. As illustrated, they have important

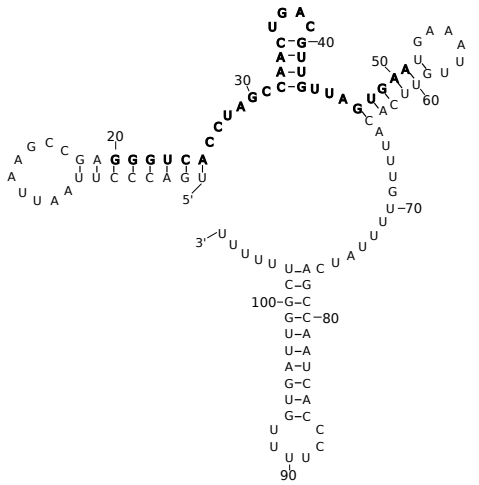


Figure 1: Predicted MFE structure for VcQrr3 with the conserved region (20 – 51 of 107 nucleotides) shown in bold. VcQrr2 has a comparable four-armed MFE prediction while VcQrr4 has an additional helix forming a "cumberbun" across the middle. VcQrr1 has the common first and last helices, but different base pairings forming a single middle arm.

commonalities as well as significant differences.

The `Sfold` [32, 29, 22, 30] approach groups structures using divisive clustering under the base pair metric [114], which counts pairings not shared between two structures. The cluster centroid, with minimum distance to all structures in the class, is the representative element. In this way, s_1 and s_2 are clustered together, with the MFE structure from Figure 1 as the centroid, obscuring critical substructural alternatives. Moreover, the similarities with s_3 are not transparent since it belongs to a second (much smaller) cluster.

Alternatively, `RNAshapes` [49, 164, 147] groups structures (by default) according to their overall branching configuration. The minimal energy structure with that shape, called a shrep, is the representative element. Both s_2 and s_3 as well as the MFE structure have the four-armed [][][][] shape, despite significant differences in the second and third arms. However, the additional "cumberbun" in s_1 gives it the [][[][]][] shape, which hides the common base pairs. Moving to a more detailed shape abstraction level helps to distinguish structural differences, but at the cost of significant similarities.

In contrast, profiling focuses on the arrangement of helices at the substructure level. Unlike methods using the base pair metric, we do not distinguish the red and purple helices in s_1 from those containing one less pairing in s_2. However, unlike branching configuration

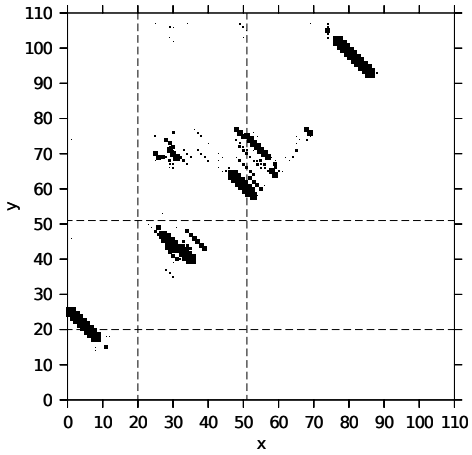


Figure 2: Dot plot of base pair probabilities for VcQrr3. Dot size at (x, y) corresponds to log probability of position x pairing with y. Dashed lines indicate the conserved region on each axis. While the first and fourth MFE helices are highly probable, the rest of the sequence — including the majority of the conserved region — has significant suboptimal structural alternatives, as well as many low-frequency pairings.

approaches, we do not abstract away all base pair details. Hence, profiling is based on a "fuzzy" definition of helix with a limited degree of elasticity in its exact composition.

We show this degree of abstraction has two benefits. It enables major structural patterns to stand out without getting overwhelmed by minor differences in stem composition. Yet, it retains enough information about specific base pairs to generate experimentally testable hypotheses.

Moreover, our method differs substantially from the existing helix-based analysis approach. Unlike profiling, `RNAHeliCes` [68, 69] does not mine the structural signal from a Boltzmann sample, nor does it classify a given set of secondary structures. Rather, their helix index shape (hishape) abstractions are generated exhaustively starting from the MFE.

These abstractions closely resemble `RNAshapes` with the refinement that helices are indexed by their "central position." Thus, the hishape of the VcQrr3 MFE structure is [13, 37.5, 55.5, 89.5] since, for instance, the first arm ends at base pair (8, 18) and $13 = \frac{8+18}{2}$. Despite this additional information, hishapes still don't characterize important relations among low-energy secondary structures.

By default, three hishapes for VcQrr3 are output. However, the MFE one still includes

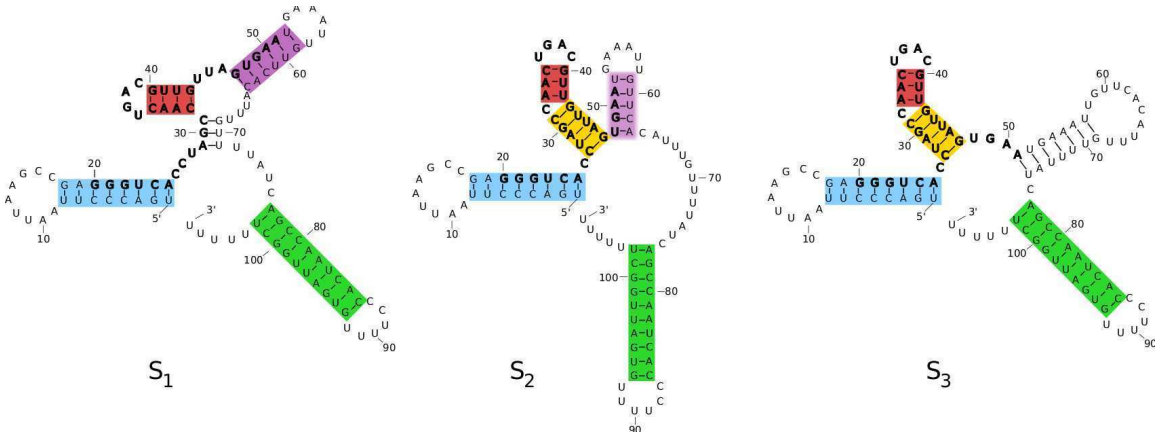


Figure 3: Three structures from a Boltzmann sample for VcQrr3 generated by `GTfold` [152] with conserved nucleotides 20 – 51 in bold. Commonalities are highlighted by colored rectangles. Significant differences include pairing 29 – 31 with 69 – 71 to form a multiloop in s_1 versus with 43 – 45 in s_2 and s_3 to form a stem extension (yellow). In s_1 and s_2, 48 – 50 are paired with 61 – 63 forming part of a hairpin stem-loop (purple) but are single-stranded in s_3.

s_2. While s_3 is now distinguished (with 63 replacing 55.5), s_1 does not appear unless additional output is requested. However, the number of different hishapes grows exponentially, with much index repetition. But since indices do not correspond uniquely to maximal helices (c.f. Figure 1 of [68]) these are not necessarily similar pairings.

In contrast, profiling identifies well-defined combinations of base pairs that dominate low-energy secondary structures with an emphasis on highlighting significant similarities and differences. This makes it well-suited for probing function, especially for regions with competing structural alternatives.

1.3 Methods

Profiling identifies and presents signal on two levels: helices and their combinations. This requires "denoising" the set of observed base pairs to highlight the dominant substructures. We employ *equivalence classes* to consolidate similar substructure elements, and *thresholds* to highlight the "head" or core of the distribution. This extracts the signal from our Boltzmann sample, yielding estimated probabilities characteristic of the entire ensemble. By truncating the low-probability tail, we retain the most frequent elements as an informative, concise, and reproducible summary of the Boltzmann ensemble.

The profiling pipeline takes a representative sample as input and outputs the substructural signal in the Boltzmann distribution. To begin, we partition the helices in our Boltzmann sample into *helix classes*. Thresholding yields the most prominent components of helix level signal, which form our set of *features*. Each structure is categorized according to its combination of features, called a *profile*. Choosing the highest frequency profiles yields *selected profiles*, whose relations are visualized in a *summary profile graph*.

1.3.1 Helix classes

Helices are a fundamental subunit in RNA structures. Under the NNTM, a secondary structure is a set of pseudoknot-free, canonical base pairs. A consecutive run of pairings $\{(i,j),(i+1,j-1),\ldots,(i+k-1,j-k+1)\}$ is grouped into a helix denoted (i,j,k). Thus, in Figure 3, $s_1 = \{(1,25,8),\ (29,71,3),\ (32,43,4),\ (47,64,6),\ (77,102,10)\}$.

When comparing secondary structures, particularly those in a Boltzmann sample, a helix in one may be a proper subset of a helix in another. For instance, the helix $(33,42,3)$ in s_2 and in s_3 is a subset of $(32,43,4)$ in s_1. At the helix level, this difference is negligible, and all three are colored red in Figure 3. Likewise with the purple helices.

Helix classes are defined to group together helices which are "the same" in this way. More precisely, a helix is *maximal* if $(i-1,j+1)$ and $(i+k,j-k)$ would be non-canonical base pairs or if $j-i-2k<5$. That is, a maximal helix respects the minimum hairpin length of 3 and is non-extendable under the Watson-Crick pairings A $\leftrightarrow$ U and C $\leftrightarrow$ G as well as the wobble pairing G $\leftrightarrow$ U.

A *helix class* consists of all helices h which are subsets of the same maximal helix g, and will be denoted $[g]$. Thus, $(33,42,3)$ and $(32,43,4)$ are elements of the set $[(32,43,4)]$, along with four other helices of minimum length ≥ 2. Given a set of secondary structures S (with multiplicity), profiling identifies the helix classes ordered by descending frequency.

The frequency of a helix h, denoted $f(h)$, is the number of times it appears in S. When S is large enough (typically of size 1000 [32]), then $f(h)/|S|$ is a good approximation to the probability of h in the Boltzmann ensemble and S is called a *representative sample*. Since $(33,42,3)$ occurs in 328 of 1000 sampled structures and $(32,43,4)$ in 573, their estimated

probabilities are 32.8% and 57.3%.

Similarly, the probability of a helix class c is approximated using its frequency $f(c)$, which is the sum over all $f(h)$ for each helix h in c. Including the frequencies of the other four helices in $[(32, 43, 4)]$, its estimated probability is 94% which is a much stronger signal than any individual helix.

1.3.2 Features

Profiling consolidates similar substructures via helix classes, thereby amplifying their signal. However, there remain many whose signal is weak at best; as illustrated in Figure 4(a), the distribution of frequencies typically has a very long tail. In this case, more than 78% of the VcQrr3 helix classes occur in less than 1% of the Boltzmann sample.

Profiling removes the "noise" of low probability pairings to highlight significant helices as our features. Hence, helix classes are selected in order of decreasing frequency, up to some threshold. In separating signal from noise, we avoid hard cut-offs, thereby substantially increasing the reproducibility of our results. Instead, profiling identifies the point of diminishing returns, where increasing the number of features begins diluting the structural signal.

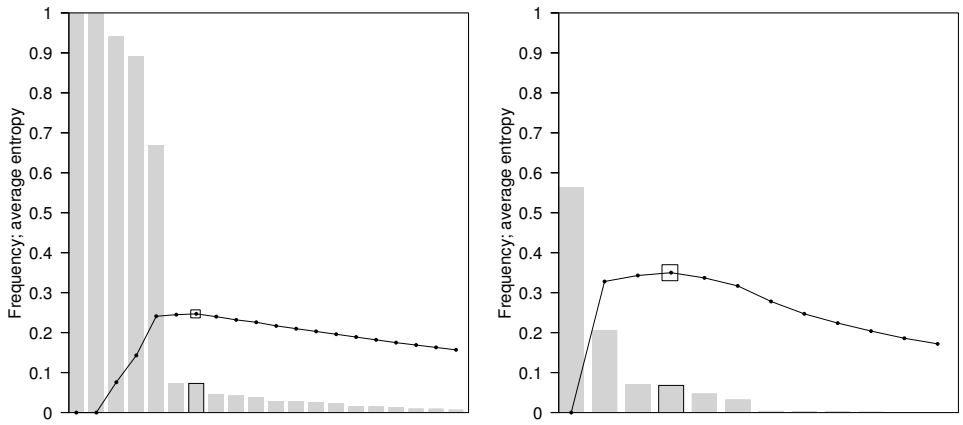
This is achieved using the concept of Shannon entropy from the mathematical theory of information. The entropy of a (binary) random variable is a measure of its uncertainty, which is also understood as information gain. The point of diminishing returns in feature selection is determined by the maximum average entropy.

More precisely, the presence of a helix class c in a structure from the Boltzmann sample is a binary random variable X_c. To ensure that the average entropy rises to a maximum, consider the estimated probability normalized by the most probable helix class c_1;

$$p(X_c) = \begin{cases} f(c)/f(c_1) & \text{if } X_c = 1 \\ 1 - f(c)/f(c_1) & \text{if } X_c = 0 \end{cases}$$

Using this rescaled probability, the entropy of X_c is calculated as

$$H(X_c) = -\sum_{x=0,1} p(x) \log\, p(x).$$



(a) Helix classes (b) Profiles

Figure 4: VcQrr3 histograms of estimated probabilities in descending order with graphs of average entropy according to Equation 1 below and its profile equivalent. The 194 helices observed in the representative sample of 1000 structures were consolidated into 88 helix classes. Only the first 20 are pictured; the estimated probability of the 20th one is 0.8%. All 13 profiles are pictured but the last seven have frequency < 5. The maximum average entropies at the 7th helix class and 4th profile are marked.

Given observed helix classes $c_1, c_2, \ldots, c_m$ ordered by decreasing frequency, we compute the average entropy at helix class c_k as

$$h_k = \frac{1}{k} \sum_{i=1}^{k} H(X_{c_i}) \text{ for each } k \text{ with } 1 \leq k \leq m. \tag{1}$$

Our threshold t is the index which maximizes this running average, and our set of *features* is then $\{c_i \mid 1 \leq i \leq t\}$.

We can prove that if there exists a k such that $H(X_{c_{k+1}}) < h_k$, then $h_i < h_k$ for all $i \geq k + 1$. Hence, if there is a local maximum h_k, then it is a global one. There are pathological distributions where the average entropy will increase until the last helix class c_m. However, for all observed distributions, the maximum occurs near the beginning of the long tail.

One advantage to thresholding by average entropy is that determining where to truncate the noisy tail is a function of the head of the distribution. Specifically, if the frequencies drop precipitously, this method will retain more low frequency helix classes than if the decline had been more gradual. In this way, lower frequency alternatives are considered only when they add value to the structural information.

Returning to our VcQrr3 example, we see this behavior illustrated in Figure 4(a), where

i	max. helix	$f(c_i)$
1	(1, 25, 8)	1000
2	(77, 102, 10)	1000
3	(32, 43, 4)	940
4	(47, 64, 7)	891
5	(27, 47, 5)	669
6	(51, 75, 7)	74
7	(29, 71, 3)	73
8	(44, 78, 3)	46

i	profile	$f(q_i)$
1	$\{c_1, c_2, c_3, c_4, c_5\}$	564
2	$\{c_1, c_2, c_3, c_4\}$	205
3	$\{c_1, c_2, c_3, c_5, c_6\}$	70
4	$\{c_1, c_2, c_3, c_4, c_7\}$	68
5	$\{c_1, c_2, c_4\}$	47

Table 1: The top eight VcQrr3 helix classes c_i (left) ordered by decreasing observed frequency. The maximum average entropy threshold is $t = 7$, so the set of features is $\{c_i \mid 1 \leq i \leq 7\}$. The top five profiles q_i (right) similarly ordered. The threshold for selected profiles is $t = 4$.

the maximum average entropy occurs at the 7th helix class — following the steep drop in frequency from the 5th one. (The first eight helix classes are given in Table 1.) Hence, our set of features is $\{c_1, \ldots, c_7\}$.

1.3.3 Profiles

Features serve two purposes. First, they highlight the core of the helix class distribution, that is the runs of base pairs which dominate the low energy secondary structures. Second, they provide the basis for understanding higher order structural signals at the combination-of-helices level.

The *profile* of a structure s is its maximal set of features. Given the set of features $\{c_1, \ldots, c_7\}$ from Table 1, the profile of the MFE structure in Figure 1 is $\{c_1, c_2, c_3, c_4\}$. This will often be denoted as (1)(3)(4)(2), using parenthetic notation with helix class indices to indicate the nesting relationships. The structures s_1, s_2 and s_3 in Figure 3 have profiles (1)(7(3)(4))(2), (1)(5(3))(4)(2), and (1)(5(3))(6)(2) resp.

Each profile is an equivalence class of secondary structures. The *specific* frequency of a profile q, denoted $f(q)$, is the size of this equivalence class, that is the number of structures in the sample S having exactly that set of features. The specific frequencies of the top five VcQrr3 profiles are given in Table 1. Note that the MFE profile is *not* the most frequent one.

We also define the *general* frequency of q as the number of structures in S whose profile

contains at least those features. Although the specific frequency of the MFE profile q_2 is only 205, its general frequency is 837 since that includes the structures from q_1 and q_4 as well.

1.3.4 Selected profiles

Like helix classes, profiles group together similar structures, thereby amplifying their signal. However, there will also be profiles with a weak signal. As before, we use a maximum average entropy threshold to truncate the distribution yielding our *selected profiles*.

The denoising calculations are essentially the same; the association of a profile q to a structure s is a binary random variable X_q. The selected frequency $f(q)$, rescaled by the most frequent profile q_1, yields a probability for the outcomes of X_q which is used to calculate the Shannon entropy. The threshold value t gives the maximum average entropy over the top t profiles, and the set of selected profiles is $\{q_1, \ldots, q_t\}$.

Figure 4(b) shows the average entropy against the estimated probability of each VcQrr3 profile. As listed in Table 1, the 1st, 3rd, and 4th selected profiles include (resp.) structures s_2, s_3, and s_1 from Figure 3 while the 2nd includes the MFE.

Selected profiles are maximal probable *combinations* of helices — a signal from the Boltzmann ensemble above the level of base pair probabilities but below whole structure groupings. As such, they are well-suited for analyzing significant similarities and differences across low-energy secondary structures. This is critical information for a molecular biologist seeking to understand which competing structural alternatives are most likely to occur in nature.

1.3.5 Summary profile graph

As illustrated in Figure 5, the relationships among selected profiles can be visualized graphically. To our knowledge, this is the first such compare/contrast summary of a Boltzmann ensemble, and should be of significant utility to researchers.

All profiles have a partial order given by set inclusion ($q \leq q'$ if $q \subseteq q'$) which is visualized as a Hasse diagram. Furthermore, the general frequency of $q \cap q'$ is at least the sum of the specific frequencies for profiles q and q'. Thus while each selected profile is a common

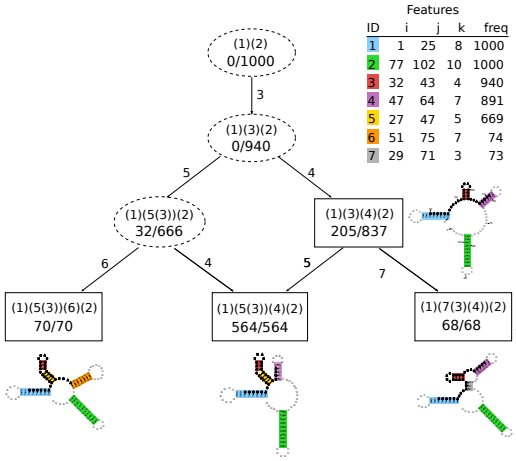


Figure 5: VcQrr3 summary profile graph. Boxes indicate selected profiles, and dashed ovals the intersection ones. Each node is labeled with the profile, in parenthetic notation, along with its specific and general frequencies, written as a ratio. An edge from q to q' is labeled with the feature(s) from $q' \setminus q$. Similarities between profiles are given by the greatest lower bound, aka "last common ancestor," with differences read from edge labels. The root is always the (possibly empty) profile common to all sampled structures. Features are listed by maximal helix with frequency. For illustrative purposes, the secondary structures from Figures 1 and 3, with features highlighted in color, are shown with their selected profile.

combination of features, their intersections are also a significant substructural signal.

To identify common substructures across selected profiles $Q = \{q_1, \ldots, q_t\}$, we calculate their intersections $I = \{q_i \cap q_j \mid 1 \leq i < j \leq t\}$. An *intersection profile* belongs to $I \setminus Q$.

We construct the *summary profile graph* using the fewest intersection profiles to (weakly) connect all selected profiles. The graph has vertices from $I \cup Q$ and directed edges between two profiles if one covers the other in the partial ordering. That is, there is an edge from q to q' if there is no q'' in $I \cup Q$ such that $q \subsetneq q'' \subsetneq q'$.

Since every sampled structure is included in at least one vertex, this graph provides a detailed yet concise overview of the most probable substructures in the Boltzmann ensemble. Reading from the top, the general frequency of the first vertex will always be the size of the Boltzmann sample. Hence, we know that every observed structure includes features c_1 and c_2, and also others since the specific frequency of (1)(2) is 0. Following the first edge, we see that 94% of the sample, and all selected profiles, also include c_3. Beyond this intersection profile, important structural alternatives begin to emerge.

Crucially, these differences all involve base pairs from the conserved region 20–51. For

instance, the region 26–31 after c_1 and before c_3 has *three* distinct possibilities: stem extension (c_5) with 66.6% probability, rare helices or single-stranded with 20.5%, or multi-branched loop (c_7) with 6.8%. The first case is read from the intersection profile (1)(5(3))(2) which includes in its general frequency two downstream selected profiles: (1)(5(3))(6)(2) and (1)(5(3))(4)(2). The second and third are the specific frequencies for the other selected profiles which include the MFE structure and s_1, resp. As will be discussed after the next section, all three cases merit further study and experimentation.

1.4 Results

As we have shown, denoising the VcQrr3 Boltzmann sample yields combinations of base pairs — features and selected profiles — which dominate the low-energy secondary structures. Moreover, as will be discussed next, the value of this substructural signal is maximized by highlighting its multimodal nature, that is the commonalities and differences which provide crucial information for molecular biologists.

First, we give proof-of-principle results that profiling successfully denoises arbitrary Boltzmann samples at this length scale. The 15 test sequences, given in Table 2, all have (1) high sample compression, so that profiling's output is a substantial reduction in scale from the input; (2) low information loss, so that features and selected profiles cover a disproportionate amount of the observed substructures; (3) reproducible results, so that variability in threshold cut-offs between independent trials is minimized; and (4) characteristic frequencies, so that the estimated probabilities extracted from the sample are a true signal from the Boltzmann ensemble. (The last case confirms that denoising via thresholding introduces no distortions in the substructural signal.)

For our test set, we selected three Qrr, tRNA, 5S ribosomal RNA, THF riboswitch, and TPP riboswitch sequences. The average length was 99nt. In our experience, the strength of the profile signal from a Boltzmann ensemble degrades significantly in the 150–200nt range. As we will explain further in our concluding remarks, this is consistent with the well-known negative correlation between MFE accuracy and sequence length [106, 35], and is the subject of ongoing research.

Although prediction accuracy is typically much higher for short sequences, there is still a wide range overall. Hence, our test sequences were arbitrarily chosen to span the range of MFE accuracies. (The Qrr sequences have unknown native structures and varying MFE predictions.) We observed little correlation with profile characteristics.

For each sequence, we generated 25 Boltzmann samples using `GTfold` [152]. Below and in the Supplementary Data, we report averages and standard deviations across samples for the same sequence, and highlight minimum/median/maximum values for comparisons among the 15 test sequences.

We find that profiling consistently identifies a small set of substructures that dominate the observed base pairing information. These results validate our VcQrr3 case study; by reducing the noise of low-frequency base pairs, profiling extracts a concise and informative substructural signal. Moreover, the thresholding of features and selected profiles is reproducible across multiple runs, and reliably characterizes the Boltzmann ensemble.

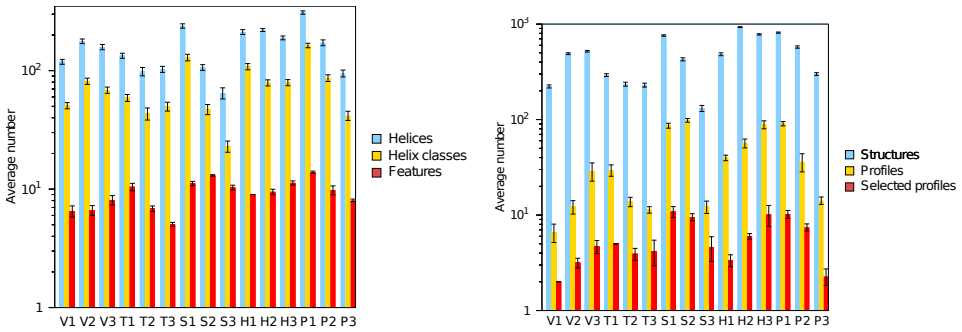
1.4.1 High sample compression

A Boltzmann sample typically contains many different helices and secondary structures. Equivalence classes and thresholds reduce the noise of low-frequency base pairs, highlighting the substructural signal presented in features and selected profiles. As seen in Figure 6, and in Supplementary Tables 1 and 2, there are a large number of unique helices on average in each sample and an even larger number of distinct structures.

Consolidating very similar substructures and truncating the low frequency tails of the distributions produces a much stronger and clearer signal. On average, the number of features and selected profiles are low enough to be investigated by hand — a substantial reduction in scale from the original sample.

We calculated compression ratios for each step and the final results. The typical noise reduction in moving from helices to features is nearly 19-fold and more than 80-fold for structures to selected profiles.

Taken together these numbers demonstrate that profiling consistently extracts a concise core of frequent substructures from a noisy Boltzmann sample.



(a) Helix compression (b) Structure compression

Figure 6: Average number of substructures in 25 samples of 1000 structures for each test sequence. Error bars indicate standard deviations. For additional clarity a log scale presentation is provided in Supplementary Figure 1.

1.4.2 Low information loss

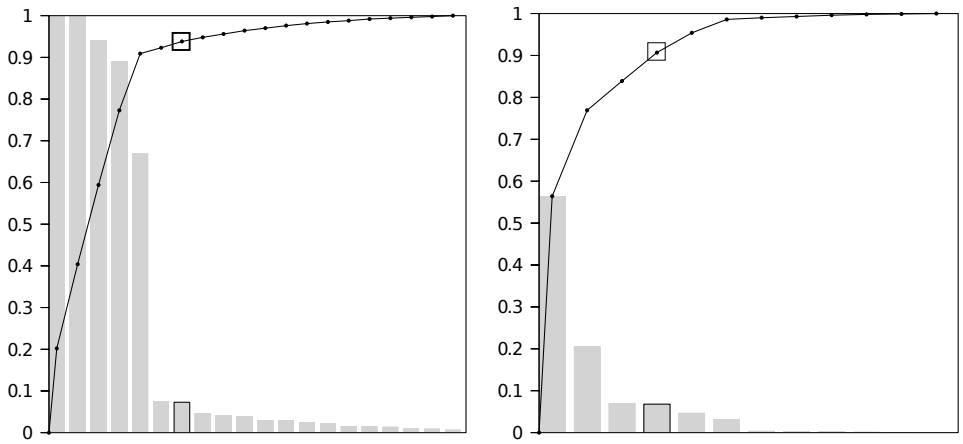
Importantly, high sample compression does *not* cost significant structural information. We measure this by calculating the coverage provided by features and by selected profiles, which is the threshold location on the cumulative density function. This is pictured in Figures 7(a) and 7(b) resp. for VcQrr3, with results for all test sequences in Supplementary Table 3.

The information loss in moving from helices to features is very low, since the typical coverage is nearly 90%. The typical selected profile coverage is nearly 83% accounting for a disproportionate amount of the observed structures. Hence, the noise reduction achieved by equivalence classes and thresholds extracts a small set of substructures which dominate the observed base pairing information.

1.4.3 Reproducible results

A significant advantage to denoising the structural signal from the Boltzmann sample is the reproducibility of profiling across multiple trials. While we certainly cannot remove all variability from this stochastic process, our results confirm a high level of stability in the occurrence of features and of selected profiles.

A feature's stability is the percentage of 25 trials in which it appears; if a helix class is above the average entropy threshold in 20 Boltzmann samples, its stability is 0.8. We calculate the feature reproducibility of a sample by averaging the stabilities of its features. Each sequence thus has an average feature reproducibility over 25 trials.



(a) Top 20 helix classes (b) All 13 profiles

Figure 7: Frequency histograms for VcQrr3 case study with superimposed cumulative distribution functions. Coverage is computed by counting the number of helices (resp. structures) with multiplicity included in the feature set (resp. selected profiles). The features cover 93.8% of observed helices (with multiplicity), and structure coverage for the selected profiles is 90.7%. Results for all test sequences are in Supplementary Table 3.

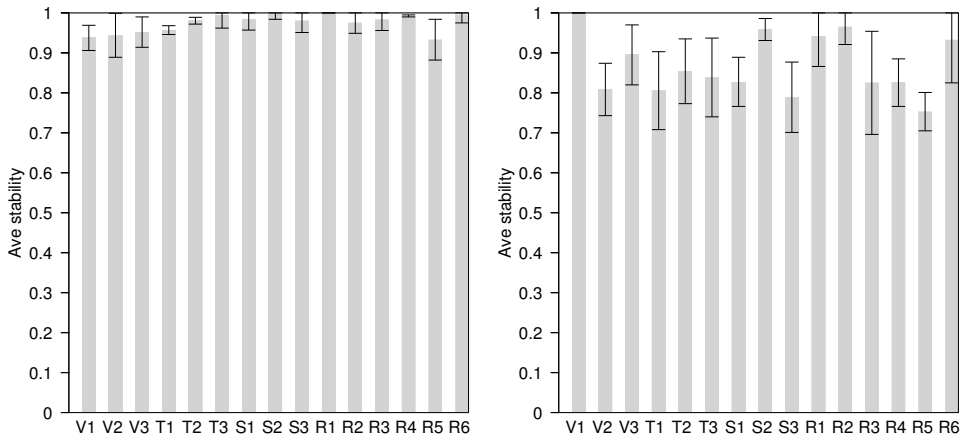
As seen in Figure 8(a) and Supplementary Table 4, average feature reproducibility is very high with minimal standard deviation for all test sequences. Hence, there is relatively little variation between sets of features across different trials.

This analysis is repeated for selected profiles. However, any differences in features will propagate to instabilities in profiles. Hence, as pictured in Figure 8(b), the selected profile reproducibility, while still high, is lower and more variable. Nonetheless, a feature or selected profile output in one trial has a high probability of being output in another.

1.4.4 Characteristic frequencies

Lastly, we confirm that profiling identifies a true substructural signal from the Boltzmann ensemble. Specifically, we measure the reliability of our helix classes and profiles by the standard deviations of their average frequencies across our 25 independent samples. The amount of acceptable variation is benchmarked by the estimated base pair probabilities [32].

For each base pair b, consider the random variable X_b whose values are the different observed frequencies of b across the 25 Boltzmann samples with equal probability. Note that if a base pair does not occur in a sample, its frequency for that trial is zero. The mean and standard deviation of X_b are then $\mu_b = E[X_b]$ and $\sigma_b = \sqrt{E[(X_b - \mu_b)^2]}$.



(a) Features (b) Selected profiles

Figure 8: Average reproducibility of features and selected profiles across 25 trials for each of 15 test sequences. Error bars indicate standard deviations.

For VcQrr3, the standard deviations for all 439 observed base pairs are visualized in Figure 9, column (a). Hence, a VcQrr3 structural signal is reliable if the maximum variation in sampled frequencies is on the order of 20 structures.

Repeating this analysis for each of the 236 helix classes observed in 25 trials gives the results in Figure 9, column (b). As expected, consolidating helices into helix classes results in a more reliable signal than individual base pairs.

Yet, there can be small fluctuations in feature selection across different Boltzmann samples. Hence, we confirm that the features from any trial yield characteristic profiles for every trial. Conditioning on a given set of features permits comparisons across all trials, and confirms that the resulting profile frequencies are reliable.

Let F be the set of features for a single Boltzmann sample, and p a profile according to F. We perform the same type of analysis for the random variable X_p across the 25 trials as for the observed base pairs. The results are given in Figure 9, columns (c) – (f) for the four feature sets observed in our 25 VcQrr3 trials. There were 12 profiles with $F = \{c_1 - c_6\}$, 15 with $\{c_1 - c_7\}$, 18 with $\{c_1 - c_6, c_8\}$, and 21 with $\{c_1 - c_8\}$. (Feature information is in Table 1.) In each case, the standard deviations for profiles are on the order of those for base pairs.

Results for the other test sequences are given in Supplementary Tables 5, 6, and 7 and

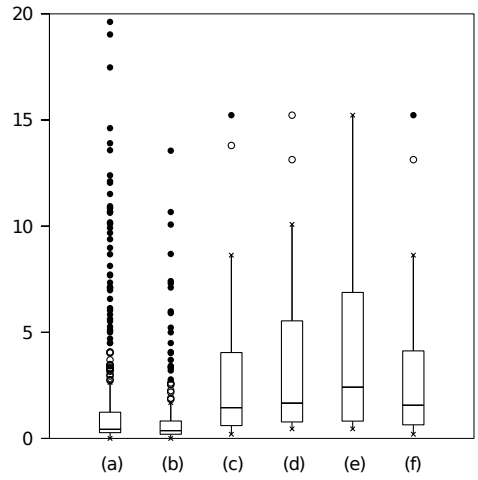


Figure 9: Box plots showing range of standard deviations in frequencies across 25 VcQrr3 Boltzmann samples. Columns correspond to (a) base pairs, (b) helix classes, and profiles conditioned on feature sets (c) $\{c_1 - c_6\}$, (d) $\{c_1 - c_7\}$, (e) $\{c_1 - c_6, c_8\}$, and (f) $\{c_1 - c_8\}$. (Features are indexed in Table 1.) Box midline indicates the median (second quartile). Top and bottom edges mark the first (Q_1) and third (Q_3) quartile, with inter-quartile range R. Whiskers indicate the furthest point within $1.5R$ of Q_1 and Q_3. Open circles are within $3R$; closed circles are beyond.

Figure 2. In all cases, the variability of the profile frequencies for a given set of features is on the order of the base pair frequency variation. Thus, in any given sample, we have confidence that the selected profiles are a true signal from the Boltzmann ensemble.

Hence, a sample of 1000 structures is sufficient for profiling to extract a clear and concise, informative, reproducible, and characteristic signal regarding significant combinations of helices for sequences at this length scale.

1.5 Discussion

We return to our VcQrr3 motivating example to discuss the benefits of profiling small RNA molecules, especially the generation of experimentally testable hypotheses.

Profiling's balance between abstraction and specificity supports and complements experimental research. By focusing on significant combinations of features, profiling highlights similarities and differences at the substructure level unhampered by sampling noise. With this information, a molecular biologist can target specific nucleotides in experiments to elucidate function.

For example, sequence alignment [100] revealed a highly conserved and likely functional region at nucleotides 20–51 in VcQrr3. According to the summary profile graph in Figure 5, six of the seven features (all but c_2) intersect this region. Both $c_1 = [(1, 25, 8)]$ and $c_3 = [(32, 43, 4)]$ have very high frequency, so the real variation occurs in subregions 26–31 and 44–51.

Nucleotides 26–31 between c_1 and c_3 have *three* distinct possibilities accounting for 94% of the sampled structures: stem extension (intersection profile (1)(5(3))(2)), single-stranded or rare helices (profile $p_2 = (1)(3)(4)(2)$), and multibranch loop (profile $p_4 = (1)(7(3)(4))(2)$). The first is the most probable (66.6%). However, the second (20.5%) includes the MFE structure which closely resembles that for VcQrr2. Furthermore, the third (6.8%) includes the analog of the VcQrr4 MFE structure. Hence, all three cases would merit further study and experimentation.

That a conserved region has a multimodal structural signal is vital information since it suggests possible functional scenarios. For instance, VcQrr3 target activation may require nucleotides 26–31 to be single-stranded. If so, these six nucleotides should have particular functional value.

By now, extensive experiments have tried to pinpoint exact mechanisms for VcQrr target interactions [58, 7]. This has involved exhaustive, systematic point mutations to verify key functional nucleotides [140, 73, 174]. Crucially, these experimental results validate the new profiling insights.

Evidence indicates base pairing with four known targets occurs in this subregion: quorum sensing response regulator LuxO at 26–33 [159], high cell density master regulator HapR at 26–45 [158], low cell density master regulator AphA at 5–30 [135, 140], and gene `vca0939` at 26–44 [58].

Furthermore, certain mutations in the 26–31 subregion knock out function in the last three cases: position 31 for HapR control [73], 25–28 for AphA activation [140], and position 28 for `vca0939` [174]. Thus, experimental evidence confirms 26–31 as especially important within the conserved region.

The profiling analysis also suggests that subregion 44–51 has a multimodal structural

signal. Nucleotides 44–46 are base paired in profiles p_1 and p_3 which contain c_5, and single-stranded or in rare helices in p_2 and p_4. Likewise, 48–50 are base paired in c_4 and not in c_6, which occur in disjoint profiles.

As with 26–31, the different possible structures have functional implications; it may be that base pairing with Qrr targets is regulated by the occurrence of different helix classes. Although this subregion has not yet been the subject of much experimental testing, the single-stranded nucleotides 58–68 in the c_6 hairpin include another region (58–65) of perfect conservation among Qrr sequences [100].

Thus, profiling identifies two critical subregions within the conserved region revealed by Qrr sequence alignment. Both have multiple different structural possibilities across the selected profiles. The importance of subregion 26–31 is validated by previous experiments, making 48–50 (as well as 58–65) a leading candidate for further investigation. It would be particularly interesting to know if VcQrr3 adopts different profile conformations *in vivo*, with major biomedical implications if virulence is deactivated in any.

1.6 Conclusion

For RNA sequences on the order of 100nt, profiling identifies dominant combinations of base pairs in low-energy secondary structures according to the NNTM. By design, this approach extracts a substructural signal from a Boltzmann sample which is clear and concise, informative, reproducible, and reliable. Moreover, by their combinatorial nature, profiles can be easily compared and contrasted, especially through the summary profile graph. Since features are tied to specific base pairs, this computational analysis generates new functional insights, facilitating experimental research such as understanding small RNAs' role in the mechanisms of cholera.

However, like all thermodynamic RNA secondary structure methods, profiling is fundamentally dependent on the NNTM's approximation of nature. In particular, it is possible to have a strong but inaccurate signal, or to have no strong signal at all, from the Boltzmann ensemble. While this is seldom an issue for short sequences, the problems become more acute as length increases [106, 35]. These issues manifest in profiling as a combinatorial

explosion of profiles for longer sequences, consistent with the exponential growth in the number of possible secondary structures [148] and abstract shapes [49]. Thus, although the feature signal remained strong in extensive testing of longer sequences, the profile signal decayed with sequence length.

Nonetheless, profiling has value beyond its demonstrated worth in analyzing small RNAs. It provides a new framework for understanding the scope and limitations of the structural signal from a Boltzmann ensemble, with potential for future enhancements. For example, the distribution of helix classes is an ensemble signature, and its stability under NNTM perturbations can be analyzed, yielding a parametric understanding of this substructure landscape. In summary, the advantages offered by profiling's combinatorial nature and balanced level of abstraction should be of significant utility to both theorists and experimentalists alike.

1.7 Availability

The `RNAStructProfiling` C code is freely available via `http://gtfold.sourceforge.net/profiling.html`.

1.8 Supplementary data

Supplementary Data are available at the end of the dissertation

1.9 Funding

The National Institutes of Health [NIGMS R01 GM083621 to C.E.H]; Burroughs Wellcome Fund [CASI #1005094 to C.E.H]. Funding for open access charge: Burroughs Wellcome Fund [CASI #1005094 to C.E.H].

1.9.0.1 Conflict of interest statement.

None declared.

1.10 Acknowledgements

We thank Dr. Shel Swenson for help with figures as well as GT students A. Panlilio and C. Mize for their transitive reduction code and D. Esposito for data processing assistance. We also thank the reviewers for their helpful feedback.

Abbr	Seq	Organism (Seq subtype)	Ref	Len	Acc
V1	Qrr	*V. cholerae* (#1)	[100]	96	–
V2	Qrr	*V. cholerae* (#3)	[100]	107	–
V3	Qrr	*Vibrio harveyi* (#1)	[158]	95	–
T1	tRNA	*Homo sapiens* (Cys)	AC004932	72	0.00
T2	tRNA	*Sulfolobu tokodaii* (Lys)	BA000023	74	0.45
T3	tRNA	*Oryza nivara* (Ala)	AP006728	73	1.00
S1	5S	*Escherichia coli*	V00336	120	0.26
S2	5S	*Acheilognathus tabira*	AB015591	120	0.59
S3	5S	*Desulfurococcu mobilis*	X07545	133	0.88
H1	THF	*Mitsuokella multacida*	ABWK02000009	99	0.11
H2	THF	*Clostridium botulinum*	CP000939	101	0.43
H3	THF	*Streptococcus uberis*	AM946015	91	0.62
P1	TPP	*Thermoplasma acidophilum*	AL445064	107	0.00
P2	TPP	*Pasteurella multocida*	AE004439	93	0.30
P3	TPP	*Bacillus clausii*	AP006627	100	0.62

Table 2: Information for 15 test sequences from five types of short RNA: Qrr, tRNA, 5S ribosomal RNA, THF riboswitch, and TPP riboswitch. Accession numbers are given for reference when available, and citations otherwise. The tRNA and 5S rRNA sequences and pseudoknot-free secondary structures were obtained from the Comparative RNA Website [21]. The THF and TPP riboswitch sequences and their consensus secondary structures were obtained from the Rfam database [55, 19]. MFE secondary structures were predicted by `GTfold` [152] using default settings. The accuracy was calculated as the F-measure, that is the harmonic mean of the MFE sensitivity and positive predictive value against true positive base pairs in the downloaded structures. Sequences were arbitrarily chosen to span the range of MFE accuracies.

CHAPTER II

NEW INSIGHTS FROM CLUSTER ANALYSIS METHODS FOR RNA SECONDARY STRUCTURE PREDICTION

This chapter is published in *Wiley Interdisciplinary Reviews: RNA* 7, no. 3 (2016): 278-294.

2.1 Abstract

A widening gap exists between the best practices for RNA secondary structure prediction developed by computational researchers and the methods used in practice by experimentalists. Minimum free energy (MFE) predictions, although broadly used, are outperformed by methods which sample from the Boltzmann distribution and data mine the results. In particular, moving beyond the single structure prediction paradigm yields substantial gains in accuracy. Furthermore, the largest improvements in accuracy and precision come from viewing secondary structures not at the base pair level but at lower granularity/higher abstraction. This suggests that random errors affecting precision and systematic ones affecting accuracy are both reduced by this "fuzzier" view of secondary structures. Thus experimentalists who are willing to adopt a more rigorous, multilayered approach to secondary structure prediction by iterating through these levels of granularity will be much better able to capture fundamental aspects of RNA base pairing.

Keywords: RNA secondary structure; minimum free energy; Boltzmann sampling; cluster analysis

2.2 Introduction

Computational methods for RNA secondary structure prediction have been an important resource for experimentalists since the early 1980's [117, 186, 184]. Prediction of a single minimum free energy (MFE) structure as the native was one of the first approaches [186, 184] and remains the most popular. MFE prediction has enjoyed this remarkable longevity due to its degree of accuracy, especially for shorter sequences [106, 35], and the simplicity of

dealing with a single structural prediction.

However, in the past three decades the RNA computational community has moved beyond the single MFE secondary structure prediction paradigm, yielding improvements to prediction accuracy [105, 108, 131]. Moreover, mounting experimental evidence indicates that many critical cellular processes are mediated by changes in RNA (secondary) structure [33, 28, 144, 90]. Hence, there are now strong biological, as well as computational, reasons for considering an ensemble of possible structures instead of just one.

In addition to new methods for generating possible secondary structures [89, 172, 34, 121, 2], significant advances have been made in refining approaches grounded in thermodyanmic optimization. Two critical enhancements to MFE predictions have included considering characteristics of individual base pairs [106, 109, 182, 74, 65, 181, 71, 180, 104] and of other low-energy alternatives to the MFE prediction known as suboptimal structures [169, 178, 171]. Importantly, these two approaches are now unified by the methodology of sampling structures from the Boltzmann distribution for a given sequence [32] according to base pair probabilities [109].

Yet, despite the demonstrated improvements in prediction accuracy from Boltzmann sampling [30, 31, 164], in practice MFE prediction programs like Mfold [179] still dominate among experimentalists[1]. The purpose of this paper is to convince the reader that this gap can and should be bridged.

The power of the Boltzmann sampling approach rests on the ability to extract key structural information from a representative set (typically of size 1000) of suboptimal structures. This is achieved by a data mining technique known as cluster analysis [79] in which similar structures are grouped together to reveal underlying patterns. Currently, there are three programs, Sfold [29], RNAShapes [147], and RNA profiling [134], which implement different approaches to secondary structure cluster analysis. The crucial differences in methodology rest on how each defines "similar" structures. This, in turn, is fundamentally a function of the granularity of the given method. Thus, in the next section, we first summarize each of

[1]According to Google Scholar, Mfold citations since 2014 are easily double the next dozen or so competitors combined.

the three methods, along with a related deterministic approach (RNAHeliCes [68]), through the lens of structural granularity.

Next, we compare and contrast these methods with each other and with the MFE prediction based on accuracy, precision, size of results, and efficiency. We show that at the level of secondary structure prediction the differences between Boltzmann clustering programs are not significant. Moreover, the representative structure for the most probable cluster for any of the three programs is at least as good as the MFE prediction. Hence, experimentalists who wish to retain the simplicity of a single structural prediction should simply replace the MFE one with the most probable representative structure to achieve better accuracy on average.

Our analysis goes well beyond this, however. We show that there are significant gains in prediction accuracy to be obtained by moving beyond the single structure paradigm. In particular, there frequently exists a representative structure with markedly better accuracy among the other probable clusters. Hence, experimentalists who view secondary structure predictions as generating a small set of possible configurations, to be vetted by further computational analysis, experimental testing, and/or biological insight, will be well-rewarded for their efforts.

Finally, we demonstrate that the largest improvements in accuracy and precision come from viewing secondary structures not at the base pair level but at lower granularity/higher abstraction. Along with a representative structure, methods which employ abstraction assign to each cluster a signature which captures the structural similarities at the chosen level of granularity. It is these signatures which truly harness the power of the Boltzmann sampling approach. Hence, experimentalists who are willing to begin with a "fuzzier" approach to understanding secondary structures will be much better able to capture fundamental aspects of RNA base pairing.

Because of the different granularity levels at which each method operates, from the fine-grained base pairs of Sfold through the higher level helices of profiling to the most abstract "topologies" of RNAshapes, these cluster analyses are not merely competitors which each other in improving over the MFE prediction. Rather, they offer complementary approaches

to representing, grouping, and comparing structures which can be used in conjunction to great advantage.

To illustrate the advantages offered by a more iterative approach to RNA structure prediction via the power of Boltzmann sampling and cluster analysis, we discuss the challenge of aptamer design. In this way, we show that ad-hoc comparisons of single MFE structure predictions can yield to a more rigorous, multilayered approach which draws on a wealth of computational advances.

2.3 Methods

While the cluster analysis methods all vary in their details, the critical difference is their level of structural granularity. The granularity used by each method informs its clustering approach, illuminates the differences between the methods, and highlights the utility of each method for different applications.

Hence, granularity is the organizing principle of this paper. In particular for this section, we describe the granularity of each method, and its ramifications for (1) structure representation, (2) clustering method, (3) representative structure, and (4) cluster signature. (Granularity also has ramifications for the type of scenario most appropriate for each method, which will be addressed in the Discussion section.) We use the example sequence given by each method to illustrate both their granularity choice and the original issue it was designed to address.

2.3.1 Sfold: at the base pair level

Having pioneered Boltzmann sampling for RNA secondary structures, Sfold was the first to tackle the challenge of data mining a set of suboptimal structures. Sfold recognized that the sample contains important information beyond the MFE structure, particularly when the native structure is not the MFE structure. One example is the *A. tumefaciens* 5S sequence, whose native conformation is markedly different from the predicted MFE (Figure 10). Accordingly, Sfold [29] identifies different viable low energy structures from a Boltzmann sample.

Sfold represents one end of the granularity spectrum by operating at the finest level of

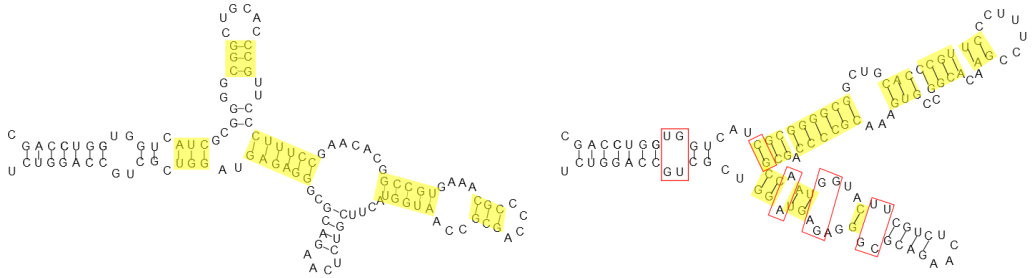


Figure 10: The two Sfold cluster centroids for *A. tumefaciens* 5S. The first is the MFE structure, the second very close to the native; they respectively represent clusters with probabilities 62.1% and 37.9%. Base pairs in the symmetric difference are shown in yellow and total 47. Base pairs separating the second from the native are shown in red; many are noncanonical. Note that single stranded bases do not count toward the symmetric difference.

resolution: the base pair level. This fine-grained approach is reflected in its representation of structures as a set of base pairs (i.e. a set of canonical pairs of nucleotides according to the allowed pairings A $\leftrightarrow$ U, C $\leftrightarrow$ G or G $\leftrightarrow$ U). Sfold also compares structures in these terms, defining the distance between two structures as the number of base pairs in either one but not in both (the symmetric difference of the two sets of base pairs).

With this well-defined metric, classic clustering algorithms can now be employed to group suboptimal structures together [22]. Sfold uses a divisive hierarchical clustering algorithm [79], beginning with all elements in a single cluster. Successive steps divide the cluster with largest diameter (maximum base pair distance between any two elements). Sfold computes twenty clusters before determining which division is optimal.

At each step, the quality of clustering is assessed with the Calinski-Harabasz (CH) index [20], a data mining metric previously used to good effect in microarray analysis [25]. The CH index calculates the ratio of distances between clusters over distances within clusters; the higher the ratio, the better the clustering. Sfold selects the clustering division between two and twenty with the highest CH index as the optimum.

These clusters capture critical information about the Boltzmann ensemble, namely that there may be more than one significant energy well present. This information is embodied in the structure chosen to represent each cluster, called the centroid structure. The centroid by definition minimizes the total base pair distance to all structures in the cluster [30].

Qualitatively, centroids reflect the high frequency base pairs of the sample, which have been shown to have higher positive predictive value (PPV) [104]. Quantitatively, centroids show improvements in sensitivity and PPV over the MFE when compared against the native [30].

This is the case with the *A. tumefaciens* 5S sequence, whose native structure is not the MFE but a low energy alternative. Thus, its Boltzmann sample yields two centroids (Figure 10), one for the MFE energy well and the other for the native one. By broadening the search beyond a single MFE structure, Sfold's analysis identifies a major structural group with almost the same frequency as the MFE cluster, and substantially more accuracy.

2.3.2 RNAshapes: at the branching pattern level

Developed around the same time as Sfold, RNAshapes operates at the other end of the granularity spectrum. While Sfold represents and clusters its structures at a base pair resolution, RNAshapes does so with respect to gross morphology. Its high level of abstraction serves as an intuitive way to cluster and manage a large number of low-energy suboptimal structures [164].

RNAshapes represents structures in terms of their topology, or *shape*, by abstracting away internal loops, bulges, and the location and length of helices. Nesting and adjacency information are preserved, and embodied in its abstract shape, as denoted by pairs of well-formed brackets. By representing structures with their abstract shapes, RNAshapes then clusters structures with the same shape together.

Each cluster has the common shape as its signature, and the number of constituent structures as its frequency. To enable structure prediction, each group is also represented by the structure with the lowest free energy of the group, known as its *shrep*.

Like Sfold's clustering, shape analysis reveals patterns in a sample about which nothing is known. Additionally, this abstraction is particularly useful when a general topology is suspected *a priori* concerning the sequence, e.g. when the sequence is related to other characterized sequences by homology or experimental data. By grouping structures with a common shape, RNAshapes enables researchers to zero in on a topology of interest [164].

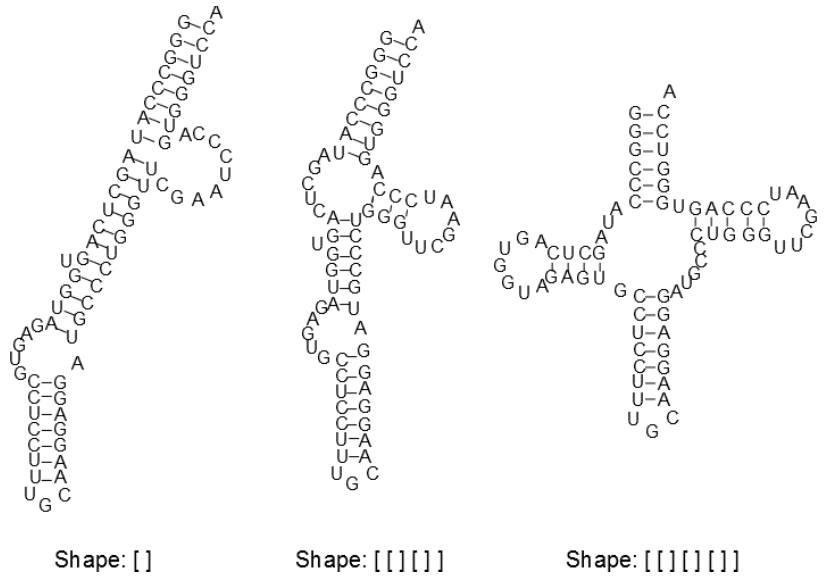


Figure 11: The three shapes present in a *N. pharaonis* tRNA-ala sample, with their *shreps*; their probabilities from left to right are 99.0%, 0.7% and 0.3%. The MFE is the *shrep* for the first, most populous shape, while the native is the *shrep* for the last.

An example of this discussed by RNAshapes and reprised here is the sequence *N. pharaonis* tRNA-ala [164], whose native structure is the well-known tRNA cloverleaf. However, the MFE has a markedly different topology of one long extended helix. Identifying low energy candidates for the native possessing the appropriate shape is difficult, without organizing structures based on topology.

RNAshapes' analysis of *N. pharaonis* yields three distinct shape groups, seen in Figure 11. The MFE structure belongs to the most frequent (incorrect) shape, which dominates the sample at a frequency of 99%. Without the benefit of shape analysis, many structures would have to be sifted through in search of one with the desired cloverleaf topology. With shape analysis, the native structure is easily located as the *shrep* of the third shape [50].

Thus, RNAshapes enables very quick perusal of the different topologies present in a set of suboptimal structures. This view of the sample at a low level of structural granularity gives one important way to summarize the structural information of the sample. This is useful when first exploring the characteristics of a sequence, but especially useful if a known topology is suspected.

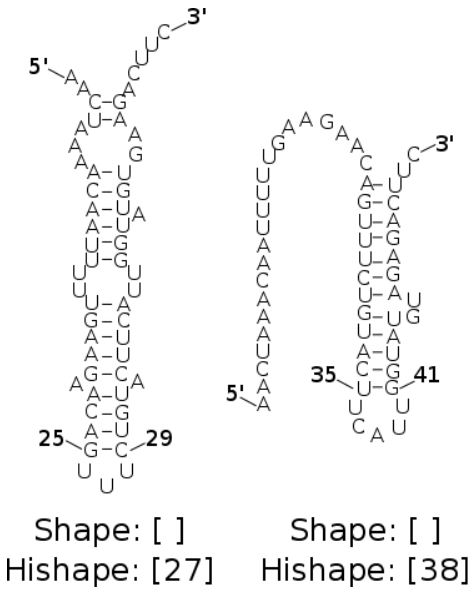


Figure 12: The two alternating native structures for the spliced leader RNA from *Leptomonas collosoma*. Both have the same shape [ ], but different *hishapes*. The first structure has the innermost base pair $(25, 29)$ and thus an index of $\frac{25+29}{2} = 27$; its *hishape* is [27]. The second structure has a helix midpoint of $38 = \frac{35+41}{2}$ and a *hishape* of [38].

2.3.3 RNAHeliCes: a refinement of RNAshapes

Developed as an extension of RNAshapes, RNAHeliCes [68] operates at a granularity between the fine gained Sfold and the abstract RNAshapes, and hence is included in this review for its interesting abstraction scheme. However, in contrast to the other methods discussed here, RNAHeliCes does not stochastically sample from the Boltzmann distribution. Rather, it deterministically enumerates all low energy structures, beginning with the lowest ones, until by default three *hishapes* are identified. While its abstraction scheme is of interest, this abbreviated analysis of suboptimal space limits its practical use. Nevertheless, we discuss RNAHeliCes for its unique granularity level, and as a general contrast to the more preferred Boltzmann sampling methods.

RNAHeliCes' intermediate level of granularity is appropriate in scenarios when multiple structures of interest have the same shape and must be differentiated. Such is the case for the spliced leader RNA from *Leptomonas collosoma* [98], their test sequence [68]. Since its two structures (seen in Figure 12) both have the shape [ ], using shape abstraction would identify at most one of them. Thus, a finer grained abstraction is needed.

Specifically, RNAHeliCes adds an index to every bracket pair in the shape abstraction to form a helix index shape, or *hishape*. The index is calculated as the average of the indices of the closing base pair and serves to differentiate helices located at different nucleotide positions, unless they are centered at the same position.

Like RNAshapes, RNAHeliCes uses abstraction as its organizing principle, clustering structures with the same *hishape* together. Allowable differences within a *hishape* group include exact helix composition, length and location of stack extensions, and internal loops and bulges. While each group has the *hishape* signature common to all its structures, it is also characterized by a representative structure (called a *hishrep*) that is the minimum free energy structure in the group.

RNAHeliCes can be used to view common *hishapes* but also to predict structure, as with *L. collosoma*. By distinguishing between stems centered around different midpoints, it identifies two *hishapes* within the common shape. For *L. collosoma*, the *shreps* for each *hishape* approximate the two alternate structures for the sequence. Thus, this level of abstraction is more appropriate to the *L. collosoma* sequence than RNAshapes.

2.3.4 Profiling: at the helix level

Like RNAHeliCes, profiling operates at an intermediate granularity, disregarding certain low frequency base pairs to consider only common helices. Developed to take a more modular approach to clustering structures, profiling enables the structural behavior of a subsequence or region of interest to be investigated [134]. Such regions include known functional domains and any new regions of interest discovered through experimental or computational means [8].

We consider the *Vibrio cholerae* quorum regulatory sequence *VcQrr3* example used by profiling [134]. No native structure is known, although a large portion of it is evolutionarily conserved with other quorum sensing sequences [100]. With sequence conservation pointing to functional and hence structural importance, the structural patterns of the given region need to be determined.

Profiling addresses this scenario by taking a helix-centric approach to representing and clustering structures. By focusing on high frequency helices known as features, profiling

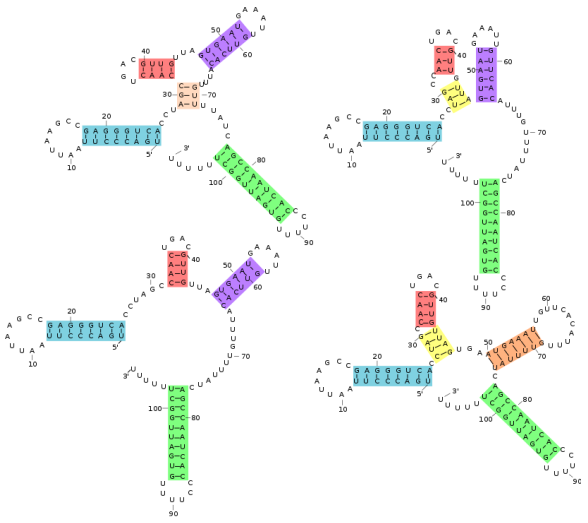


Figure 13: Four VcQrr3 consensus structures, with colors indicating different features. Their probabilities are, clockwise from top left, 6.8%, 56.4%, 7.0% and 20.5% Each structure as a combination of colors illustrates profiling's representation of a structure as a set of features. The MFE structure is the lower left.

represents structures by their particular combination of features, known as its profile. Structures with the same profile are clustered together, and the most frequent profiles are selected as clusters of particular interest. The clusters with their profile signatures thus summarize the helical information in a Boltzmann sample. By abstracting away low frequency helices, common patterns involving the key helices and thus key regions can emerge.

In addition to highlighting regional and helical trends, profiling can be used for simple structure prediction. Thus in addition to a signature profile, a representative structure is given for each cluster. This is the consensus structure, which is composed of all the base pairs present in a majority of structures in the profile. Profiles provide a more abstract way of viewing the salient information in a cluster, while the consensus structures give a base-pair resolution view of the information.

The consensus structures for the four *VcQrr3* selected profiles are seen in Figure 13. Previously, these profiles were shown to contain four distinct structural patterns in the conserved region [134], with the variations centering around nucleotides known as key to functionality [140, 73, 174]. Thus, for extracting information at the regional level, profiling clears away lower level details and presents major helical patterns.

2.4 Evaluation criteria

As illustrated, each cluster analysis method extracts important information from sets of suboptimal structures at complementary levels of granularity. In addition to these individual proof-of-principle results, in this section we compare the methods in four key measures: accuracy, precision, result size and runtime.

Accuracy is always a factor when choosing a method, as is the practical issue of runtime. Precision, or the repeatability of results, is an issue due to the stochasticity of Boltzmann sampling. Finally, since we are moving beyond considering just one MFE structure to multiple suboptimal ones, the typical number of clusters returned is an important characteristic of the analysis method.

As described, the methods associate both a representative structure and a signature to each cluster. To investigate the effect of abstraction, we evaluate the performance of the signatures as well as the more commonly assessed structures. It will be shown that although representative structures across methods have comparable accuracy, increased abstraction typically yields increased accuracy as well as increased precision.

We evaluate these measures using ten Rfam [55] families with sequence length less than 200, the range of best performance for thermodynamic optimization [35]. From each Rfam family alignment, ten seed sequences were chosen to give a median and average MFE F-measure score of approximately 0.5. For the accuracy comparisons, the native base pairings for each sequence were obtained by aligning it with the Rfam consensus structure.

For the computations, we use GTfold, a parallel implementation of the MFE algorithm [152], Sfold 2.2, RNAshapes 2.1.6, RNAHeLiCes 2.0.14 and profiling 1.0.

2.4.1 Accuracy

There are several options when measuring accuracy, the first and most important assessment of a method. Unlike the MFE optimization, the cluster analysis methods return multiple clusters, each represented by both a structure and a signature. Thus, there is the option of measuring the accuracy of one structure or of the aggregate of multiple ones, and of doing so for signatures as well. As we shall see, there are reasons for exploring all these options.

As others have done [30], we report the accuracy of the best representative structure, i.e. the structure with the highest accuracy. This gives a sense of the best the methods can do, and of the fundamental limitations to accuracy each method is bound by. However, when the native is not known, the best structure cannot be identified. Instead, the highest frequency or most probable structure, is always apparent; hence, we also calculate its accuracy.

Moreover, in addition to considering the accuracy of a single structure, we will argue that considering multiple structures is worth the improvement to accuracy. As some researchers may be able to systematically investigate all structures, we calculate the overall average accuracy of all structures, both unweighted and weighted by the frequency of the cluster. Comparing unweighted versus weighted indicates the general frequency of more accurate structures; if accurate structures are of lower frequency, then the unweighted accuracy will be greater, and vice versa.

Representative structures, however, are not the only structural information produced about the clusters. These methods also give information at higher abstraction levels in the form of cluster signatures. We will show that structural predictions on a broader scale than base pairs have a better accuracy than the representative structure, and hence are also evaluated in addition to the more traditional structure level.

Table 3: Information for the ten test families, each having ten test sequences. MFE accuracies are calculated with F-measures using the GTmfe package of GTfold and the native structure from the Rfam consensus alignment. The median score is reported in the table. Sequence length reflect average family length as reported by Rfam, which were used in selecting the ten families.

ID	Description	Length	MFE acc.
UnaL2	UnaL2 LINE 3' element	54.1	0.59
tRNA	transfer RNA	73.4	0.51
Intron group II	Group II catalytic intron	87.2	0.56
THF	THF riboswitch	99.6	0.51
TPP	TPP riboswitch	111.6	0.5
5S	5S ribosomal RNA	116.6	0.53
U5	U5 spliceosomal RNA	117.2	0.52
FMN	FMN riboswitch	136.6	0.52
U1	U1 spliceosomal RNA	162	0.53
U2	U2 spliceosomal RNA	190	0.57

For each method including the MFE prediction, we calculate its accuracy as the F-measure, which is the harmonic mean of positive predictive value and sensitivity. We summarize results for each family by reporting the median most probable, best, average and weighted accuracy over all sequences in the family.

More precisely, positive predictive value is calculated as

$$PPV = \frac{TP}{TP + FP}$$

and sensitivity as

$$S = \frac{TP}{TP + FN}$$

where TP denotes a true positive, FP a false positive, and FN a false negative. The F-measure is defined as

$$F = 2 \cdot \frac{PPV \cdot S}{PPV + S}$$

.

We compare the base pairs of the native against the predicted structure to determine accuracy. Base pairs common to both structures are counted as true positives; base pairs occurring only in the native but not the predicted as false negatives; and base pairs only in the predicted but not the native as false positives.

A more general definition of true positive, false positive and false negative involving edit distance is needed to calculate signature accuracy. The edit distance details the transformation of the native into the predicted by a series of either insertions or deletions. Any insertions to the native signature is considered a false positive, and any deletion a false negative. Any element of the native signature not deleted in the edit distance is a true positive. The shortest edit distance gives us the necessary terms to calculate the F-measure. Recall that Sfold's clusters have the centroid as both signature as well as representative structure.

For profiling, each group has its profile (a set of features) as its signature. We calculate the F-measure of selected profiles against the profile representation of the native structure, with helices that are not features omitted from the profile by definition. Common features are true positives, features found only in the profile representation of the native are false negatives, and features found only in the selected profile are false positives. For simplicity

we consider only features of length greater than two base pairs. Because very low frequency profiles are not selected, the weighted accuracies of the selected profiles are calculated using normalized frequencies.

The RNAHeliCes signature is its *hishape*. Each hishape is a set of indices with associated loop type. To calculate accuracy, we use their convert function to translate the native structure into a hishape, comparing it against the predicted hishapes with their tree edit program. A true positive is a loop type and index found in both the native and predicted hishape, with false positives and negatives found only in the predicted or native respectively. Because RNAHeliCes gives free energies and not frequencies for its hishapes, we approximate *hishape* frequency in calculating the weighted average accuracy. An abbreviated partition function from the given free energies is used as a normalizing factor to determine the probability of each hishape.

We use a similar tree edit approach to calculating the accuracy of RNAshapes signatures. For simplicity and consistency, we use the RNAHeliCes tree edit program to determine the edit distance between the native and predicted shapes. Both the native and the representative structure are translated into hishapes by the RNAHeliCes convert function before being input into the tree edit program. The additional index of *hishape* is disregarded by ignoring any relabeling edits. Insertions and deletions as before provide counts for true positives, false positives and false negatives.

2.4.2 Precision

The deterministic MFE and RNAHeliCes algorithms always return the same result for a given sequence. However, there is no such guarantee with methods analyzing a stochastically generated Boltzmann sample. Thus, not only the accuracy but also the precision (or repeatability) of results is an issue.

We measure precision by running each stochastic method ten times. The precision score for each cluster representation (structure or signature) is the observed fraction of runs in which it represents a cluster. For example, if a structure appears as an Sfold centroid in eight runs out of ten, it receives a precision score of 0.8. The precision of the cluster

representatives can thus be calculated for all methods.

Like accuracy, we report precision in four ways: the score of the most frequent element, of the best element, of the average of all the elements, and of the average of all the elements weighted by their frequencies. The most probable element is always apparent and can be used if only one element is desired, while the best element demonstrates the advantages of using multiple structures. Finally, comparing the weighted with the unweighted average reveals that precision increases when the higher frequency elements are weighted accordingly.

2.4.3 Size of results

Although accuracy results will demonstrate the viability of using cluster analysis methods even when only one structure is processed, results will also show that there is almost always a more accurate representative structure. Results size thus quantifies how many representative structures are produced. This result, in combination with others, demonstrates that using only a handful more structures pays a significant dividend in accuracy.

For Sfold, we report the number of clusters; for profiling, the number of selected profiles; and for RNAshapes, the number of shapes. We show results for RNAHeliCes as a reference only, as the default setting for this deterministic method always produces the three lowest energy hishapes.

2.4.4 Runtimes

The expediency of computational prediction methods is a significant motivator for their use. Accordingly, we quantify the efficiency of each method by its runtime.

By now, MFE methods are well-optimized, resulting in efficient runtimes. Although these cluster analysis methods have been developed more recently, we show that their run times do not suffer much in comparison. We use the runtime of GTfold [152], a parallelized implementation of the MFE method, for comparison. We measure the time it takes to generate and analyze a Boltzmann sample for a given sequence using a high resolution timer. We report the median run time across all sequences in a family.

2.5 Results

Results confirm the superiority of using cluster analysis methods instead of the MFE prediction. First, if only one possible structure will be considered, the most probable structure should be used since its accuracy is often better, and unlikely to be worse, than the MFE prediction. Second, considering just a few alternative structures confers real improvements in accuracy, so researchers are strongly urged to broaden their methodology beyond single-structure predictions. Finally, the most significant gains in accuracy, as well as precision, are achieved by viewing structures more abstractly as cluster signatures. This suggests that random errors affecting precision and systematic ones affecting accuracy are both reduced by this "fuzzier" view of secondary structures. As discussed in the next section, the implication for researchers is that the most accurate structure predictions will be achieved by iterating through the levels of granularity. Furthermore, this benefit will be maximized by coupling the computational analyses with experimental hypothesis testing.

2.5.1 Accuracy

Results confirm the preferred approach of Boltzmann sampling. Because Sfold, profiling and RNAshapes summarize the structural information from a larger, more representative group of structures, accuracy results as a whole are more reliable. Boltzmann sampling methods in general either perform at or above the level of MFE structures, while RNAHeliCes can dip significantly below (Figure 14d). Furthermore, while Boltzmann signatures (e.g. Figure 14b and Figure 14d) perform reliably better than their associated representative structures (e.g. Figure 14a and Figure 14c), this relation is not seen in RNAHeliCes. Thus, while we consider RNAHeliCes for its unique abstraction scheme, we focus primarily on the three Boltzmann sampling methods.

Within the sampling methods, the best accuracies achieved by Sfold, profiling and RNAshapes for their representative structures are close to each other (Figure 14c). No one method sustains a clear advantage, with all methods producing the best accuracy for at least one RNA family. Similarly, the top accuracy score among most probable structures does not uniformly belong to one method, but shifts between methods depending on RNA

family (Figure 14a). Thus at the base pair level of accuracy, there is little difference between these three methods. Consequently, we now compare all the methods' representative structures collectively against the MFE structure.

Figure 14a illustrates that using the most probable structure is a better strategy than using the MFE. For each method, only RNAHeliCes had one family (TPP) with accuracy below 95% of the MFE accuracy. Moreover, on average, the accuracy is usually 6% above.

However, in nearly all the cases, the accuracy of the most probable representative (Figure 14a) structure is not the best (Figure 14c). Every method has a representative structure with accuracy better than the MFE, for every RNA family. Even adding only two additional suboptimal structures for RNAHeliCes, which always produces just three structures by default, significantly improves the best accuracy in a substantial number of cases. (The improved scores of the most probable and best structures have previously been shown with Sfold [30], but we demonstrate that these results are not tied to Sfold's methodology but are a general result of clustering suboptimal structures.) Thus, while considering only one structure is the simplest, expanding the scope of investigation even a little carries significant benefits.

If resources allow for only a few more suboptimal structures to be processed, then the higher frequency ones should be considered first. This is implied by comparing the unweighted average accuracies (Figure 14e) against the weighted average (Figure 14g). For Sfold and RNAshapes, the weighted accuracy is higher than the unweighted because the very low frequency structures are unlikely to be the native pairings. In contrast, profiling already removes these structures from consideration. Hence, the lower frequency selected profiles are exactly those shown to be more accurate by the other two methods. Accordingly, for profiling the unweighted has higher accuracy than the weighted. Thus, if only a few but not all of the structures can be considered, selecting the more frequent ones is the best strategy.

Although the methods are largely interchangeable at the base pair level, this is not the case as abstraction is introduced. In a majority of the cases (e.g. Figure 14c vs. Figure 14d), the signature has a higher accuracy than its representative structure, indicating that

broader structural predictions are more correct than specific ones.

Additionally, the degree of abstraction is related with the degree of accuracy. Especially for the best and averaged accuracies (Figures 14d and 14f), shapes is clearly better than profiles, which is clearly better than centroids. The improvement in accuracy is especially significant for RNAshapes; the most probable shape is the correct one in most families (Figure 14b). This agrees with an intuitive sense that computational prediction, while not completely accurate in all the base pair details, nevertheless is sophisticated enough to predict the broad outlines of structure correctly at this length scale.

Thus, accuracy results confirm the superiority of using structures from a Boltzmann sample, preferably more than one. They also confirm the strategy of using abstraction when possible.

2.5.2 Precision

Precision increases as abstraction increases. Because a lack of precision often indicates the presence of random errors, this indicates that there is significant stochastic noise at the base pair level in Boltzmann sampling. The best scores (Figure 15c) indicate that despite stochastic noise, the Boltzmann sample has a clear signal that the methods are consistently picking up. The precision of the most probable structure (Figure 15a) is usually among the best scores of each run (Figure 15c). Hence the more frequent elements are consistently present in runs with high repeatability, with stochastic noise affecting the low frequency elements more. This is further confirmed by precision scores significantly increasing when average scores (Figure 15e) are weighted according to their frequencies (Figure 15g). Thus, considering the most probable structure, and preferably two to five other high frequency structures, as the native is advantageous not only with respect to accuracy but also to precision.

The stochastic noise is further reduced by lowering the granularity from structures to signatures. Both profiling and RNAshapes have their best and most probable precision scores boosted to perfect repeatability for all families when considering signatures (Figures 15d and 15b). Even between signatures, there is a clear inverse relation between level of

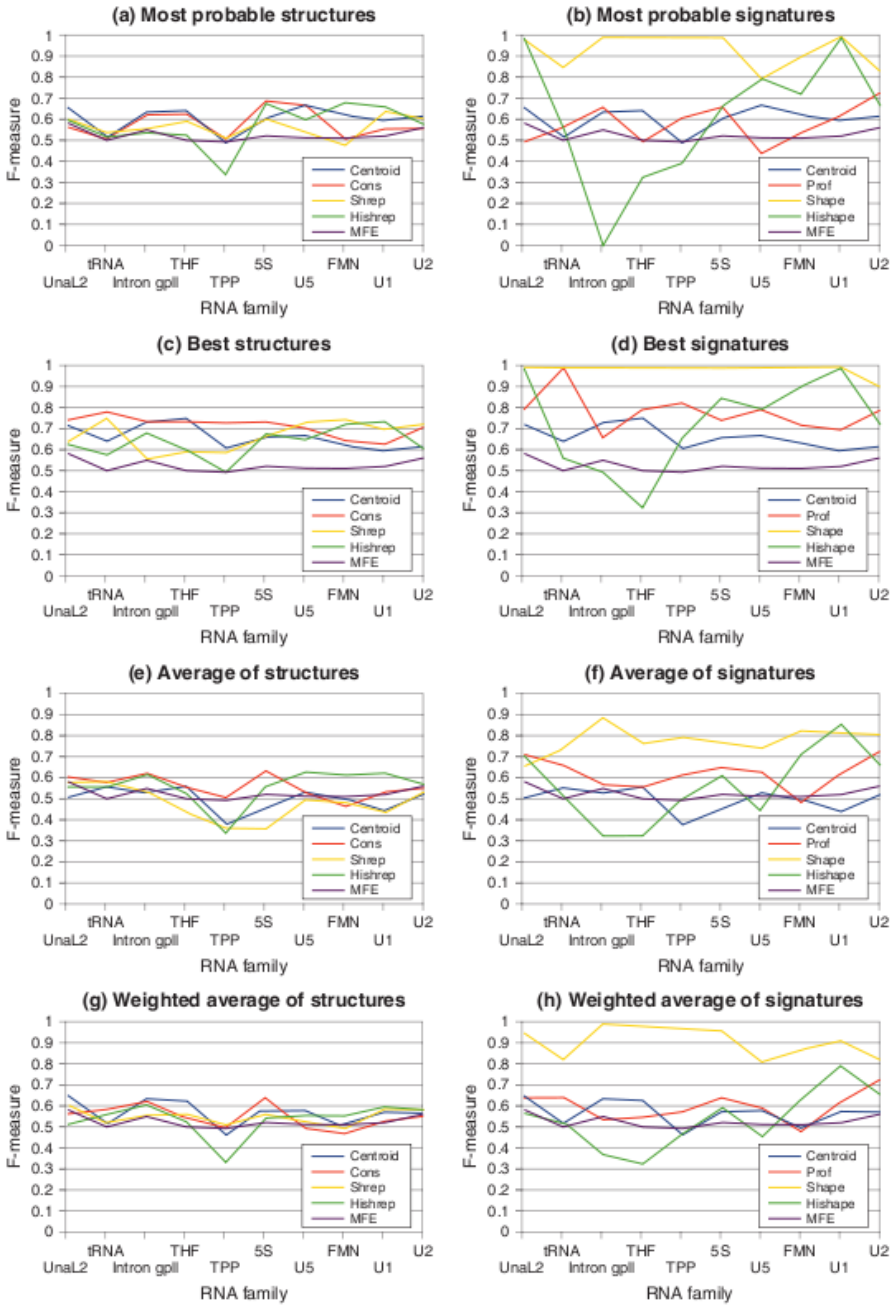


Figure 14: Accuracy comparisons for representative structures (left) and signatures (right). Median scores are reported for each family. Sfold centroids are used for both. The median MFE F-measure is also reported for comparison. Note the significant improvement in accuracy for signatures versus structures.

granularity and precision. According to the average precision scores (Figure 15f), Sfold's fine-grained centroids perform worse than profiling's more abstract helix centric view, which in turn is worse than RNAshapes' more abstract shapes. This again is consistent with accuracy results, which strongly encourage the use of signatures under the principle that the lower granularity, the better.

Taken together with the accuracy scores, we see that both accuracy and precision typically increase with higher frequency structures, and even more so with signatures. This indicates that abstraction alleviates both random stochastic errors that affect precision, and potentially more systematic ones affecting accuracy. Precision results also confirm the strategy of always considering the most probable structures, additional high frequency structures when resources allow, and abstract signatures when feasible.

2.5.3 Size of results

For researchers partial to the one structure simplicity of the MFE method, any of the Boltzmann methods' most probable structure is a better choice than the MFE. However, analyzing additional high frequency structures pays dividends in accuracy, as seen by the fact that the most probable structure is usually not the most accurate. Results size confirm that the number of additional structures to be analyzed is typically small.

Sfold consistently gives some of the smallest number of clusters, i.e. between two to six clusters. For Sfold, the number of clusters does not noticeably differ as sequence length increases. Thus, the best accuracies of Sfold are accessible by considering only a handful of additional structures, which covers all the clusters.

The median number of selected profiles is slightly larger but always under a dozen, and generally correlated to sequence length. Considering all the selected profiles is still feasible, as is focusing on the most frequent two to five profiles.

RNAshapes reports a median number of shapes ranging from two to 19, with the number of shapes increasing more significantly with longer sequence lengths. However, the growing number of shapes is populated in large part by very low frequency clusters, which have been shown to have relatively poor accuracy and precision. Hence, using a few of the most

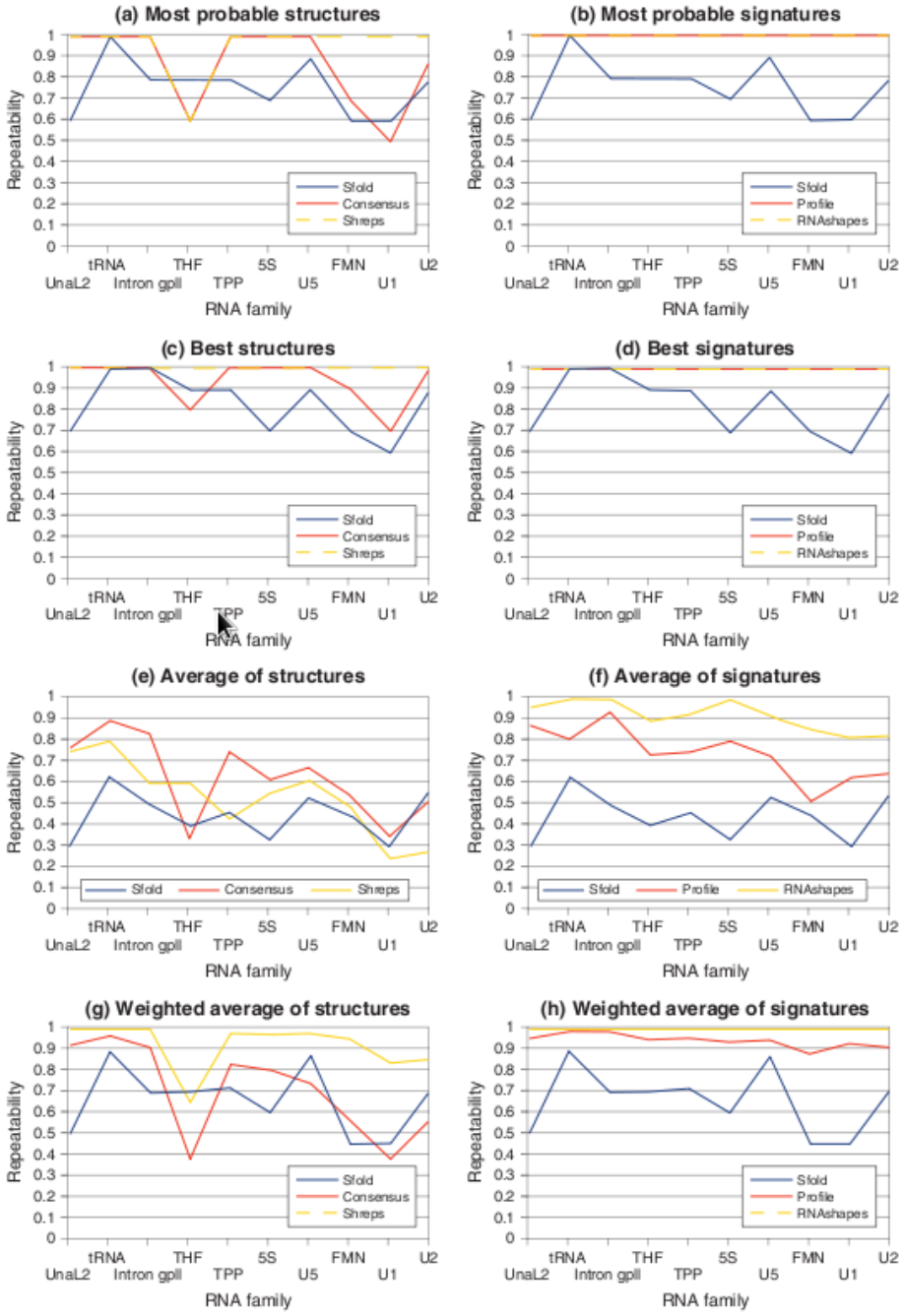


Figure 15: Precision comparisons for representative structures (left) and signatures (right). Median scores are reported for each family. Sfold centroids are used for both. Neither RNAHeliCes nor the MFE prediction are included, since both are deterministic with perfect precision. Note the improvement in precision for signatures versus structures.

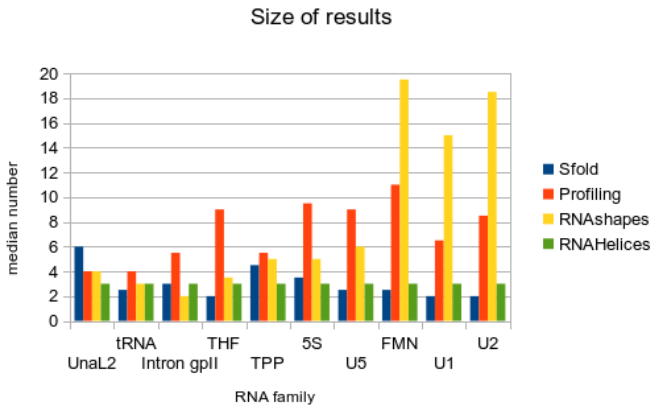


Figure 16: Median number of groups for each RNA family. RNAHeliCes always by design returns three groups, and is included here for reference.

frequent structures is again encouraged. If signatures are used, employing just the most probable is a valid strategy, given the most probable signatures' very high accuracy and perfect precision.

2.5.4 Runtime

Compared to GTfold, the cluster analysis methods are slower, though to human perception there is little difference between the runtimes of GTfold, profiling and RNAHeliCes. Hence, runtime is not a discriminating factor under normal conditions (e.g. no massive number of runs).

Sfold was fairly consistent in generating and analyzing Boltzmann samples, averaging around 25 seconds, with the most time spent in its computationally intensive clustering algorithm. Both profiling and RNAHeliCes ran in usually under a second, though RNAHeliCes' run time went up for longer families. RNAshapes' run time was inbetween Sfold and profiling, and was the most variable. Run time increased with sequence length for all methods, as expected.

Thus, while high volume studies may preclude using slower methods, single runs can be made by any method in reasonable time.

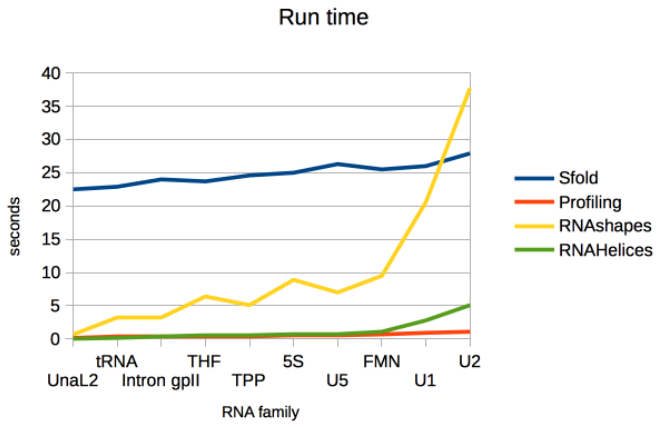


Figure 17: Median run time of Sfold, profiling, RNAshapes and RNAHelices.

2.6 Discussion

Results demonstrate that at the base pair level, the Boltzmann cluster analysis methods are indistinguishable, most notably in terms of accuracy. Hence, whether selecting one structure to use or employing the more preferred multiple structures, any of the methods is sufficient to show improvements over the MFE prediction. Real differences, however, appear when considering signatures with their differing granularity levels. Specifically, lowered granularity translates to higher accuracy and precision, indicating that errors both systematic and random are addressed at least in part by structural abstraction.

These results taken together present a clear strategy for employing cluster analysis methods: use the most probable structure instead of the MFE prediction, consider multiple structures when resources allow, and begin with signatures instead of structures when feasible. While the methods are largely indistinguishable at the base pair level, careful consideration is needed if abstract signatures are used, as each method operates at a different granularity level.

Use of the appropriate method yields information at the given granularity, which in turn should motivate further investigation. By iterating between computation and experimentation, the granularity of exploration can progress from very low to very high. Thus, while their representative structures make these methods competitors at the base pair level, their

signatures make them complementary tools from a granularity perspective.

If nothing but the broadest knowledge concerning a sequence is known, then the broadest and most abstract method (RNAshapes) is the place to begin. For example, a group of related sequences (identified through evolutionary homology, sequence alignment, or experimental results) may need to be characterized. Such a scenario occurs with aptamers, which are sequences that bind to a ligand of interest and are typically of length 100 nucleotides or less. Of increasing interest in therapeutic use [116, 82], aptamers can be found experimentally from a large random pool of sequences [70, 150, 10]. The nature of aptamer-ligand binding, however, is not well understood, nor is it always clear what the key similarity is that causes a group of sequences to all bind to the same ligand [66].

The secondary structure of the sequence is thought to be crucial to its binding, and structural features common to all the sequences are of high interest. Since little is known about the sequence(s) of interest, a very high level, shapes-oriented approach is a good starting point to identifying common branching motifs. Sampled shapes are highly accurate and precise at this level, and can be directly compared between sequences of differing lengths. If a branching pattern of interest is identified (such as the linear or slightly branched topologies known to be favored [138, 93, 48]), then only aptamers with the predicted branching pattern, for example, can be included in a experimental selection pool. This early weeding out of potentially unviable sequences could alleviate the low yield and high cost of aptamer synthesis [10], thus increasing the effectiveness of experimentation. Using shape predictions could also preclude the not uncommon scenario of generating random sequence pools with low structural diversity [48] or with characteristics different from functional molecules.

Results from aptamer selection usually produce a smaller subset of sequences with the desired binding affinity. Thus, while all sequences may have the same branching configuration, a higher granularity level is now needed to investigate details that enable some sequences to bind while others do not. The wide gap between shape and shrep accuracies in Figures 14c and 14d indicate that much accuracy is lost by jumping from a shape to the MFE structure with that shape.

By considering different helix combinations with the same shape, the focus can be narrowed further without moving directly to base pairs. Helix-centric methods like profiling give the location and length of helices within a topology, enabling the search for common regional motifs and domains, like the bulge-hairpin [10] or the stem-loop motif [26] known to be functionally important in many characterized aptamers.

Regional analysis afforded by profiling is needed when computational or experimental data points to a particular area of interest. Computationally, sequence alignment tools can determine that a conserved subsequence exists. Experimentally, subregions of interest are found in sequences from partially structured libraries [26, 118]. Shown to improve aptamer selection, these sequences typically contain a conserved subsequence flanked by two randomized subsequences. High performing sequences require regional analysis to determine the structural behavior of their randomized subsequences. Profiling a sequence gives the major combinations of helices possible for the region, enabling common motifs to emerge.

A proposed motif can be verified by screening additional sequences predicted to form similar substructures in the key region. Once a motif or domain is identified as the potential key to binding, granularity can be increased to a nucleotide level. Mutation experiments predicted to disrupt key domains can verify computational predictions or necessitate alternate hypotheses. Successful knockout mutations pinpoint specific nucleotides of interest, which can be tracked by the cluster analysis methods' representative structures. Sfold in particular can process a set of structures with only a few key base pair differences among them. Because mutagenesis experiments at this level are resource-intensive, such a fine-grained level of analysis should only be performed after iterating through coarser grained signature analysis.

Researchers can thus iterate between experimentation and computation, using one to inform the other. By employing a more nuanced use of these complementary signatures, brute force experiments can give way to faster and more informed techniques. Furthermore, ad-hoc comparisons of structures can yield to a more rigorous, multilayered approach that draws on a wealth of computational research.

Finally, given the benefits of this multilayered approach, it would be interesting to

incorporate other abstractions, such as trees [96, 143, 1] or graphs [44, 141, 84, 85] into the cluster analysis of Boltzman samples. Expanding the number and granularity of tools can only strengthen the computational benefits to experimentalists.

2.7 Conclusion

The RNA computational community has long known the advantages of considering information in addition to the MFE prediction, investigating the use of base pairs and suboptimal structures to improve accuracy. Yet, the single MFE prediction paradigm still dominates among experimentalists, although increasing biological evidence indicates that multiple secondary structures have functional significance in nature. The purpose of this paper has been to convince the reader that this gap can and should be bridged.

First, the gap should be bridged because cluster analysis of Boltzmann samples outperforms MFE predictions, even at sequence lengths where thermodynamic optimization is the most accurate. To begin, picking the representative structure associated with the highest frequency cluster from any Boltzmann sampling method is more accurate on average than the MFE. Moreover, whenever additional information (experimental, computational or otherwise) is available to discriminate between potential alternatives, multiple structures should be considered, since an even more accurate structure can almost always be found. Furthermore, the more accurate structures are likely to be the higher frequency ones, so low frequency structures need be considered only when resources allow. This, in conjunction with the relatively small numbers of clusters, typically a dozen or less, ensure that only a handful of additional structures need to be processed to improve accuracy.

Second, the gap should be bridged because the cluster analysis methods offer a more powerful function than mere structure prediction. Namely, these methods also represent clusters with abstract signatures. The signatures' different granularities provide alternative ways to view and compare structures that confer better accuracy and precision. These improved results imply that signatures help reduce systematic errors (potentially present in the thermodynamic model) and random ones introduced by stochastic sampling.

Signatures also provide complementary ways of mining the important structural information of a Boltzmann sample. These include the trends and motifs in the sample concerning branching, helical and base pair patterns. The appropriate level of cluster analysis depends on the level of information known or desired concerning a structure, i.e. very broad or general hypotheses are well suited for RNAshapes analysis and testing, more specific regional ones for profiling, and very specific base pair ones for Sfold.

The different granularity levels also indicate the viability of iterating back and forth between computation and experimentation. Computation helps guide experimentation, which generates more fine-grained hypotheses to be verified by higher granularity methods, and so forth. Because lower granularity signatures are in general more accurate, employing this graduated approach to analysis should funnel researchers toward more accurate results than leaping straight to the base pair resolution of an MFE structure.

As shown, cluster analysis methods have much to offer the experimental community, from a superior single structure prediction strategy to a more sophisticated one of iterating between computation and experimentation. These methods reflect the wealth of research relevant to real world problems developed in the last decades to turn RNA structural data into actionable information. Their adoption by the experimental RNA community will only improve current analysis and speed up the rate of important discovery and applications.

2.8 Funding

Funding for this paper was provided in part by the National Institutes of Health [NIGMS R01 GM 083621 to C.E.H]; Burroughs Wellcome Fund [CASI #1005094 to C.E.H].

2.8.1 Conflict of interest statement

None declared.

CHAPTER III

FURTHER CONSIDERATIONS TO IMPROVE PROFILING FOR LONGER SEQUENCES

3.1 Intro

The previous chapters demonstrate the success of RNA profiling on extracting structural signal from a Boltzmann sample of 1000 structures. However, one major caveat remains: signal at the profile level devolves into a messy situation for sequences appreciably longer than 300 nucleotides, such as 16S rRNA.

For these longer sequences, the signal at the feature level remains intelligible, with profiling returning a sizeable list of features that are present in the sample with high frequency. However, there are so many of features that their exact combinations in the sample result in a combinatorial explosion; there are often almost as many combinations of these features as structures in the sample. Thus, it is meaningless to profile structures with the set of features. Signal consolidation thus fails at the profile level.

To alleviate this issue, a way must be found to further condense the set of features. This is possible without much loss of structural information based on the observation that many features are very similar to another feature in size and location. These similar features therefore can be combined under a further layer of abstraction with minimal loss of information.

To identify features that can thus be combined, we determine relations between features known as *logicals*. This concept is borrowed from Boolean logicals, i.e. `AND`, `OR` and `NOT` relations. Two features that always occur in a structure together has an `AND` relation, while two that never occur together has a `NOT` relation. If the present of one implies the other, then a directional `OR` relation exists.

Profiling thus can be modified to select features in two parts: first selecting a core set of features based on frequency (original method), then augmenting this set with helix classes

Logical	Mutual info
$5 \leftrightarrow 7$	0.527
$5 \leftrightarrow 8$	0.333
$3 \leftrightarrow 9$	0.224
$4 \leftarrow 7$	0.169
$12 \leftarrow 14$	0.147
$3 \leftrightarrow 12$	0.120
$\vdots$	$\vdots$

Table 4: VcQrr3 logicals ordered by descending mutual information

that participate in relations with high mutual information. In this way we select helix classes with an eye toward satisfying the stability and information characteristics as the two criteria of signal. We select them as features, and subsequently use these features as as basic building blocks to describe higher levels of structure. Preliminary results show that both features and their combinations are more stable than helix classes or stems, with a greater concentration of data.

3.2 Augmenting features

Frequency is not the only criteria for being informative; a helix class in a relation with high mutual information also tells us much concerning the sample. Thus, we augment the core features with helix classes that partner with core features in a logicals. While these augmenting features are more likely to be unstable, they also provide key information regarding the relations of features. They are provided in addition to the core entropic features to be used at the discretion of the user, with a clear understanding that they usually lower overall stability.

Given a logical $l = (c_1, c_2)$, let $H(l) = H(c_1, c_2)$. We order L by $H(l)$ in descending order and label each logical $l_1, \ldots, l_{|L|}$, such that $H[l_1] \geq \ldots \geq H[l_{|L|}] > 0$.

We define a threshold logical $l_t = (c, d)$ such that $c, d \notin F_e$, and $\forall (c', d') \in L$ such that $H((c', d')) \geq H(l_t)$, $c', d' \in F_e$. Then $F_a = \{d \notin F_e : \exists c \in F_e \text{ s.t. } H((c, d)) > H(l_t)\}$. This threshold selects the highest mutual information logical involving two non-core features.

Table 4 shows the logicals of VcQrr3 in descending order. Recall that the threshold for core features is set at 8. The logical $12 \leftarrow 14$ (in yellow) is chosen as the threshold,

because it is the logical with the highest mutual information that does not involve two core features. Helix class 9 is an augmenting feature, as it participates in a logical above this threshold. Although other thresholds may be equally reasonable from an information perspective, results show from a stability perspective this method of choosing a threshold to be remarkably stable, unlike hard thresholds based on an absolute value. Since both stability and coverage are important, we choose this threshold method over other, less reliable methods.

3.3 Logical graph

We use the feature set $F = F_c \cup F_a$ as basic building blocks whose combinations provide a convenient way to describe structure. Namely, in deciding on a set of key helix classes, we now present their relations. We define the set of logicals defined by F to be

$$K = \{(c, d) \mid c, d \in F, (c, d) \text{ is a logical}\}.$$

We visualize this set by a directed graph $G = (V, E)$, with $V = F$ and $E = \{(c, d) \mid (c, d) \in K\}$. We use blunt double headed arrows if $c \uparrow d$.

Profiling outputs this graph after redundant relations have been eliminated; relations are deemed redundant if they can be inferred from the application of propositional logic to other relations. The logical graph output for VcQrr3 is found in Figure 18.

3.3.0.1 Biconditional relation

Consider two features c, d with $c \leftrightarrow d$. In propositional logic, We merge them into one vertex if all the relations involving c can be inferred from $c \leftrightarrow d$ and the relations of d, and vice versa. In other words, c, d are mergeable if $\forall e$ such that $(c, e) \in K$, $\{d, e\} \in K$; moreover, (c, e) is in the same relation as (d, e).

If two features in a biconditional relation do not have the same relations to a set of features, then we visualize the relation with a higher mutual information score. Namely, a difference in relations implies one feature participates in at least one relation the other does not. Without loss of generality, suppose $\exists e$ such that $(c, e) \in K$, but $(d, e) \notin K$, or (d,e)

Helix	Triplet	Frequency
1	77 102 10	999
2	32 43 4	923
3	1 25 8	780
4	27 47 5	692
5	47 64 7	565
6	2 107 2	477
7	23 50 4	391
8	51 75 7	295
9	19 28 3	124

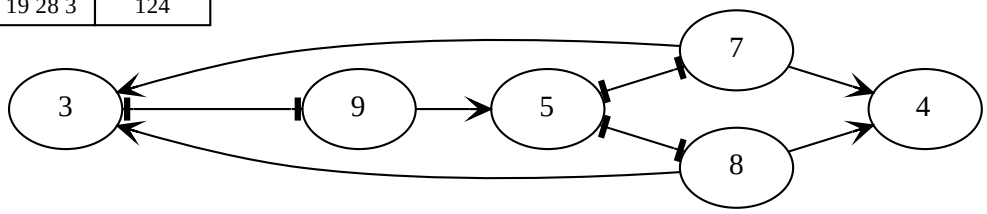


Figure 18: Logical graph for VcQrr3

is in a different relation. If $H(c,d) \geq H(c,e)$, then we merge c,d and do not display (c,e); else, we do not merge c,d and visualize (c,e).

3.3.0.2 Conditional relation

If we consider biconditional and conditional relations as conventional directed edges, then we may reduce the number of conditionals by taking the transitive reduction of the graph.

3.3.0.3 NAND relation

If a NAND relation may be inferred from other NAND and conditional relations, then it is redundant and should be eliminated. Consider two features c,d such that $(c,d) \in\uparrow$. Let $R(c) = \{d \mid (c,d) \in\rightarrow \cup \leftrightarrow\}$; this is the set of vertices reachable through the directed, conditional edges from c. Then if $\exists c' \in R(c), d' \in R(d)$ such that $(c',d') \in\uparrow$, then the edge (c,d) is redundant. This follows from the implications $c \rightarrow c'$, $c' \rightarrow \neg d'$, $\neg d' \rightarrow \neg d$. This last implication is the contrapositive of $d \rightarrow d'$. Hence, $(c,d) \in\uparrow$ is implied, and can be eliminated.

3.3.1 Subprofiles

We partition the logical graph G into its k connected components: $V = V_1 \cup \ldots \cup V_k$, with $V_i \subseteq F$. Each connected component is a set of features such that the inclusion of one in a structure affects the inclusion of the others; hence, it represents interconnected regions of the sequence.

There are many subsets of V_i that do not violate any of the relations embodied in the edges of V_i; we call any such subset a valid 'resolution' of V_i. However, a possible resolution of V_i may not be probable. In finding probable resolutions of a connected component, we identify signal at the level of common combinations of features.

Figure 18 for VcQrr3 shows that the logical graph is one connected component, with $V_1 = \{3, 4, 5, 7, 8, 9\}$. Both subprofiles $B_1 = \{3, 4, 7, 8\}$ and $B_2 = \{4, 5, 9\}$ are valid resolutions of V_1 that do not contradict any of the relations in Fig. 18. However, we shall see that a frequency may be assigned to each subprofile; B_1 has a frequency of 243, versus B_2 with a frequency of 26. As before, we focus on the more frequent elements, with B_1 of more interest than B_2.

Let the subprofile of a structure s be $B(s, V_i) = \{[h] : h \in s, [h] \in V_i\}$. The set of structures with the same subprofile b under V_i is $S_b = \{s \in S : B(s, V_i) = b\}$. To every subprofile b, we may assign a frequency $f(b) = \sum_{s \in S_b} f(s)$. Then the set of all subprofiles for V_i is $\mathcal{B}(V_i) = \{B(s, V_i) : s \in S\}$, with total structures involved $T_i = \sum_{b \in \mathcal{B}(V_i)} f(b)$.

Profiling presents subprofiles that are more likely to be stable by being more frequent. Thus, we select a threshold to filter out less frequent subprofiles. However, the subprofiles seen should at least provide structural context for every feature $c \in V_i$. We thus define a threshold taking our need for both stability and coverage into account. We consider all subprofiles that contain a helix class c, and and take the set of their frequencies: $D(c) = \{f(b) : b \in \mathcal{B}(V_i), c \in b\}$. Note that every c is uniquely associated with a particular V_i and hence a particular $\mathcal{B}(V_i)$. For all c that compose V_i, we take the maximum frequency subprofiles that contains it: $E_i = \{\max(D(c)) : \forall c \in V_i\}$. We then take the minimum such maximum subprofile c as our threshold: $b_t = c$ such that $f(c) = \min(E)$.

Table 5 shows the VcQrr3 subprofiles for Fig. 18, ordering by descending frequency.

Subprofile	Frequency
3 4 7 8	243
3 4 5	213
3 5	168
3 4 7	116
5 9	95
5	33
⋮	⋮

Table 5: VcQrr3 subprofiles ordered by descending frequency; inclusive threshold shown in yellow

Subprofile $\{5, 9\}$ is the inclusive threshold, as every feature from Fig. 18 appears in at least one subprofile with frequency equal to or above $f(\{5, 9\}) = 95$. Structures a, b, c in Fig. 3 has subprofiles of $\{3, 5\}$, $\{3, 4, 5\}$, and $\{3, 4, 8\}$ respectively. While structures a, b have subprofiles that are above the threshold and are thus reported, structure c does not; its subprofile $\{3, 4, 8\}$ has a frequency of 23.

Profiling thus reports every subprofile with frequency equal to or above this subprofile: $\mathcal{C}(V_i) = \{b \in \mathcal{B}(V_i) : f(b) \geq f(b_t)\}$. This guarantees that each helix class is seen at least once but not necessarily more, ensuring less stable subprofiles are not represented.

CHAPTER IV

CONDITIONING AND ROBUSTNESS OF BOLTZMANN SAMPLING OF RNA SECONDARY STRUCTURES UNDER THERMODYNAMIC PARAMETER PERTURBATIONS

4.1 Abstract

Understanding how RNA secondary structure prediction methods depend on the underlying nearest neighbor thermodynamic model remains a fundamental challenge in the field. Minimum free energy predictions are known to be "ill-conditioned" in that small changes to the thermodynamic model can result in significantly different optimal structures. Hence, the best practice is now to sample from the Boltzmann distribution, which generates a set of suboptimal structures. While the structural signal of this Boltzmann sample is known to be robust to stochastic noise, the conditioning and robustness under thermodynamic perturbations have yet to be addressed. We present here a mathematically rigorous model for conditioning inspired by numerical analysis, and also a biologically inspired definition for robustness under thermodynamic perturbation. We demonstrate the strong correlation between conditioning and robustness, and use its tight relationship to define quantitative thresholds for well- versus ill-conditioning. These resulting thresholds demonstrate that the majority of the sequences are at least sample robust, which verifies the assumption of sampling's improved conditioning over the minimum free energy prediction. Furthermore, because we find no correlation between conditioning and MFE accuracy, the presence of both well- and ill-conditioned sequences indicates the continued need for both thermodynamic model refinements and alternate RNA structure prediction methods beyond the physics-based ones.

4.2 Introduction

Improving secondary structure predictions remains a fundamental challenge in RNA structural modeling and design [163, 105, 183]. Thermodynamic optimization methods have

been the dominant approach for decades [108, 131, 179, 32, 64], although the problem of predicting a minimum free energy (MFE) secondary structure under the nearest neighbor thermodynamic model (NNTM) has long been characterized as ill-conditioned [177, 94]. This is usually understood as a large number of structurally-distinct suboptimal configurations within a small energy range of the MFE value [184, 105, 171], and can be successfully addressed by stochastic sampling (typically a set of 1000 structures) from the Boltzmann ensemble [32].

Equivalently, though, the ill-conditioning of RNA thermodynamic predictions can be understood as sensitivity to small changes to the NNTM [94, 35]. This is significant because the NNTM is a large objective function, with many parameters of varying degrees of precision [137, 106, 161, 166]. While Boltzmann sampling is designed to address the ill-conditioning of the MFE prediction, no studies have considered the effect of NNTM perturbations on the Boltzmann ensemble itself. This paper fills that knowledge gap by addressing two questions: (1) How well conditioned is Boltzmann sampling as a mathematical optimization problem? (2) How robust is it as a model of a biological system? We provide a rigorous quantitative answer to the first by computing the relative condition number, and answer the second by defining robustness as the persistance of a structural signal in the Boltzmann ensemble. We then demonstrate the strong correlation between this mathematically-defined conditioning and biologically-inspired robustness, and explore its major implications.

Previous work has focused on the effect of parameter perturbation on MFE structures [95, 94]. While not investigating ill-conditioning explicitly, an early study establishes a model for finding MFE structures under a normally distributed parameter perturbation [95]. More recent work took this model and used it to explicitly address ill-conditioning [94]. Results found that even slight perturbations were enough to alter the MFE structure significantly, as measured by a normalized tree metric.

We build on these previous works to further quantify and investigate both conditioning and robustness, with an increase in the scope, rigor and complexity of the analysis. To

investigate conditioning, we use the numerical analysis definition of an ill-conditioned problem as "one with the property that some small perturbation of x leads to a large change in $f(x)$" [156]. By carefully defining the change in input and change in output, we develop a novel metric not only to measure differences between samples, but also to quantify ill-conditioning itself based on established mathematical principles.

To investigate robustness, we use a biological definition of a robust system as "the persistance of a system's characteristic behavior under perturbation or conditions of uncertainty" [149]. Although robustness studies usually take the sequence as input and perturb it through simulated mutations [168, 165, 136], here we fix the sequence and perturb the NNTM to determine robustness against parameter uncertainty. We determine whether the sample under perturbation is fundamentally, structurally different (non-robust), or merely changes by the reweighting of the frequencies of the same structural elements (sample robust).

Hence, our investigation of both conditioning and robustness hinges on measuring the change in the sample under perturbation. However, because normal stochastic effects produce mild changes between Boltzmann samples even under unperturbed conditions, the measured change under perturbation should ignore these slight fluctuations. Previous work has demonstrated that high frequency pairings are more stable against stochastic fluctuations than low frequency ones [134]; hence, the former should be considered the 'signal' of the sample, whereas the latter can be considered the 'noise'. Thus, we build upon this work by tracking only the changes to the important structural signal of the sample, as represented by high frequency helices.

All possible changes affecting this high frequency signal can be partitioned into three categories defined by the scope of the frequency changes: signal that remains signal, signal that remains part of the original, unperturbed sample (though not part of the signal anymore), and signal that under perturbation ventures outside the sample into the universe of structures. These three categories correspond to decreasing levels of robustness— signal robustness, sample robustness, and non-robustness— and will be shown to be highly correlated with conditioning. This equivalence will further provide a guide for interpreting

conditioning by yielding well- and ill-conditioning thresholds. By employing these thresholds, we demonstrate that most sequences are largely sample robust, even under significant NNTM perturbation. Furthermore, because robustness is not correlated with MFE accuracy, the existence of both well- and ill-conditioned sequences point to the need for research in both NNTM refinement and complementary non-physics based prediction methods.

4.3 Methods

Our quantitative analysis is based on established principles from numerical analysis, a branch of mathematics interested in the behavior of computations under perturbation. In particular, we will compute the relative condition number, denoted κ. This is the ratio of the largest relative change in output over the relative change in input. We consider the relativized version [54, 61] since the size of the output can vary significantly over the problem instantiation. Hence, comparisons are made with the appropriate normalization.

More precisely, the relative condition number is defined as

$$\kappa = \sup_{\delta x} \frac{\|\delta f\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|}.$$

Given a function f defined for an input x and perturbed by a small amount δx, the change in output is defined as

$$\delta f = f(x + \delta x) - f(x).$$

The function is considered ill-conditioned when the (normalized) ratio of the size of these changes is large. Thus, to adapt the methods of numerical analysis, we must rigorously define x, δx, f and δf, and their respective sizes, in order to compute κ.

4.3.1 Defining the input x, the change in input δx, and their sizes

At a high level, we define the *input* to be the nearest neighbor thermodynamic model (NNTM). Its *size* is its L_1 norm when the model is viewed as a vector, e.g. $\|x\| = \sum_i |x_i|$, where each coordinate x_i is one of the thousands of parameters of the NNTM. The *change in input* is defined as 5, 10 or 20% of each parameter value. The *size* of the change in input is the L_1 norm of the change in input when viewed as a vector. We shall see that defining

these terms in this way is both simple and intuitive, leading to a clean ratio that becomes the denominator of our conditioning metric.

The name of the NNTM refers to its basic premise that the thermodynamic score of a structural component (e.g. stacked base pair or internal loop) is a function of the number and type of its nearest neighboring flanking base pairs. Thus, there are 21 parameters for the stacked base pairs, since there are six canonical base pairings but a 5'-3' symmetry to the stacks. However, the number of parameters for the different loop types is considerably higher since the composition of the adjacent single-stranded bases now also plays a role. Hence, there are almost 250 parameters for loops of arbitrary size, and over 8000 for the special cases of small internal loops. (See [161] for extensive documentation on the model.)

To obtain δx, we perturb each parameter by adding or subtracting a given percentage d of its value. For each model parameter, the direction (up or down) of the perturbation is chosen independently at random, with the amount of perturbation set to the given percentage (d = 0.05, 0.10 or 0.20) of that parameter. Although there are many known dependencies in the parameter derivations, we choose to utilize this simpler model in this initial study. (We note, though, that substructures with 5' - 3' symmetries, such as base pair stacks, are identified with a single thermodynamic parameter, and all duplicate instances in the code are perturbed consistently.)

To calculate the *size* of the input change, we consider δx to be a vector of values $\{dx_i\}$ (where x_i is the ith parameter of the model) and apply the same L_1 norm, that is, the sum of the magnitude of its values. This gives $\|\delta x\| = \sum_i |dx_i| = d\sum_i |x_i|$. When the NNTM is perturbed in this way, the relative change in input under the L_1 norm simplifies to $\frac{\|\delta x\|}{\|x\|} = d$. Perturbing by a percentage is both mathematically very tractable as well as biologically consistent, since the NNTM parameters vary in size over the different categories of substructures.

We test three values of d—5%, 10% and 20%— that are representative of the range of observed error margins [176]. Since we are interested in the worst case scenario, we generate ten sets of perturbed parameters for every d. For each d, the same ten parameter sets are used for all sequences to normalize results. Thus for every sequence, we calculate ten κ's

by iterating through the ten parameter sets, and select the highest ratio as the overall κ for the sequence.

4.3.2 Defining the output $f(x)$, the change in output δf, and the sizes of both

At a high level, we define the *output* to be the high frequency helices, shown to be the 'signal' of a sample [134]. Its *size* is the number of helices being tracked from the original, unperturbed sample. The *change in output* is the differences the signal undergoes from the unperturbed baseline sample to the perturbed sample. Its *size* is the sum of all the differences when discretized into bins of standard deviation. We shall see that tracking changes in this way captures key differences between the signals of the unperturbed versus perturbed samples, while filtering out low level differences from stochastic noise. This also enables us to track not only the magnitude of the changes for conditioning metrics, but also its source for robustness calculations.

To avoid tracking stochastic noise, we define the output $f(x)$ to be a Boltzmann sample's characteristic signal. Previous work has demonstrated that by first grouping helices into equivalence classes called helix classes, and then focusing on the high frequency ones, the signal can be isolated from the stochastic noise [134]. Hence, we define both the output f and the change in output δf in terms of high frequency helix classes.

More specifically, we have previously defined an equivalence relation on helices in order to abstract away low level base pairing differences [134]. Namely, all helices consisting of a subset of the base pairs of the same non-extendable maximal helix are placed in the same equivalence class, called a helix class. We thus consider helices to be equivalent, for example, that have the same starting and ending coordinates (i, j), differing in only in the length k of the stack. This difference, commonly seen in both stochastic sampling and in molecular dynamics, is rarely considered a significant change; this view is thus codified by these helical equivalence classes known as helix classes.

The helical signal of the sample is further concentrated by focusing on the high frequency helix classes. This is possible because every helix class can be assigned a frequency, based on the number of structures containing a member of that class. Thus, a helix class with high

frequency denotes a high number of structures possessing an equivalent helix. (A more in depth definition and explanation of these terms and results can be found in [134].) Hence, we build upon previous work by utilizing the signal, or the high frequency helix classes, as the output $f(x)$ in order to focus on key changes.

Since we have defined the output $f(x)$ to be the base pairing signal given by the high frequency helix classes, then its norm $\|f(x)\|$ is the number of helix classes in this signal. For simplicity we define all helix classes with frequency at least 10% to be the signal. (We note that the motivating results used a more nuanced, sequence-specific methodology to define the signal to avoid the stochastic instabilities inherent in a hard cut-off [134]. Here, though, we can use a simple threshold criteria because the sampling fluctuations will be addressed through our novel method for measuring the change in the structural signal δf and its size $\|\delta f\|$.)

Calculating the change in output $\delta f = f(x) - f(x - \delta x)$ should capture the meaningful differences in the structural signals between two samples. This difference both encompasses the symmetric set difference between the signals, and any significant difference in frequencies between helix classes present in both. The challenge is to do this in a way which is not sensitive to the noise from stochastic sampling; even when the NNTM is kept constant, Boltzmann sampling will produce helix classes frequencies that differ slightly. Thus, when tallying perturbation changes, we need to avoid attributing these normal frequency changes to ill-conditioning. Our approach is motivated by the understanding that values in Gaussian samples which are more than three standard deviations from the mean are significant.

Thus, in order to determine the threshold for significance, we form a model for helix class frequency in order to calculate a standard deviation σ for each one. We then use σ to filter out sampling stochasticity, and also to capture the degree of change by tallying the frequency difference in units of 3σ. Specifically, frequencies within 3σ of the mean are counted as zero, between 3σ and 6σ as one unit of change, between 6σ and 9σ as two units, etc.

In order to determine the boundaries for normal frequency fluctuations, we first model the occurrence of a helix class in a structure as a Bernoulli trial, with probability p of

success, i.e. there are pn structures containing a member of that helix class out of a sample size of n. We then can model a helix class's frequency as binomially distributed, which calculates variance as $\sigma^2 = np(1-p)$, standard deviation $\sigma = \sqrt{np(1-p)}$ and the mean $\mu = np$. Hence, as long as we have an accurate probability p, we also can obtain a reliable mean μ and standard deviation σ; any frequencies more than 3σ away from μ can then be ascribed to perturbation effects and not to ordinary sampling stochasticity.

In measuring the change under perturbation, we first obtain an unperturbed sample u, then a perturbed one b for comparison. In order to obtain a reliable p, we use a high resolution unperturbed sample of $n_u = 100,000$ structures to ensure accurate calculations of σ and μ. We denote the number of times a helix class appears in the unperturbed sample as q_u (ranging from 1 to n_u), and in the perturbed sample as q_b (ranging from 1 to n_b). We can then use $p = \frac{q_u}{n_u}$ and the more typical perturbed sample size $n_b = 1,000$ to calculate our final σ and μ. Finally, we measure the total degree of change for helix class i as $\Delta_i = \left\lfloor \frac{|\mu_i - q_b|}{3\sigma_i} \right\rfloor$. We handle any new helix classes that were previousy not present in the original sample by setting their original frequency q_u to zero; as will be explained later, because of pseudocounts their standard deviation is set to one.

Empirical tests show a good agreement between the model and observed standard deviations (Figure 19). While there are some differences in the mid-range frequencies, the agreement is solid enough, especially at the low and high frequency ranges, to use it as a valid theoretical approximation.

At high n, the binomial distribution is well approximated by a normal distribution, under which 99.7% of values lie within 3σ of the mean. Hence, fluctuations in helix class frequency occurring 3σ away from the mean are almost certainly due to NNTM perturbation. Conversely, any fluctuation within 3σ of the original mean should be ignored as indistinguishable from normal stochastic variations.

To avoid zero values of σ, which occur with helix classes of 100% frequency, we add a pseudocount to every σ. The simplest pseudocount method is Laplace's rule, commonly used in bioinformatics [41], to augment each σ by one. Hence, helix classes of 100% frequency are assigned a standard deviation of 1.

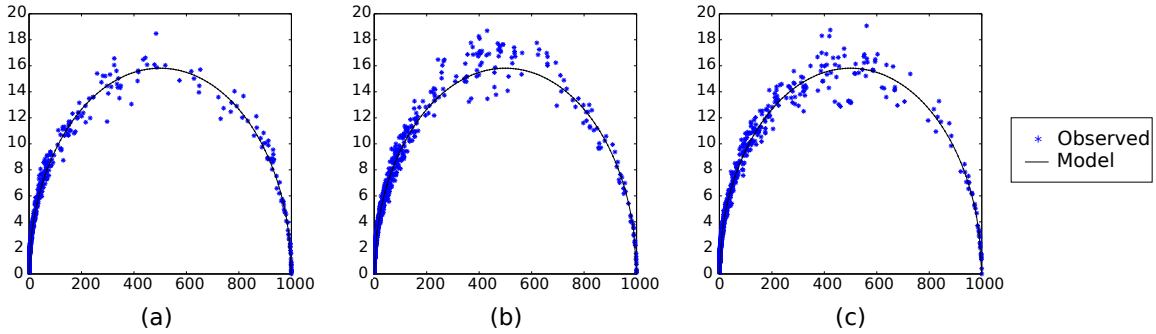


Figure 19: Actual versus model standard deviation for helix classes of (a) *H.volcanii*, (b) *E.coli* and (c) *E.cuniculi* 16S rRNA sequences. These sequences have been shown to have very different MFE accuracies and behaviors under SHAPE perturbation [151]; their helix class frequency behaviors, however, are seen to be similar, and thus are assumed to be typical. A hundred samples of 1,000 structures each were generated for the sequences, using the same unperturbed, original set of parameters. In order to gauge the normal level of helix class frequency variation, the standard deviation for each helix class frequency was calculated (i.e. the square root of the average of the squared deviations from the mean). Dots represent a helix class, with the mean μ of its frequency across 100 samples as its x-coordinate, and the calculated standard deviation σ' of its frequency across 100 samples as its y-coordinate. The curve represents the model standard deviation, calculated as $\sigma = \sqrt{np(1-p)}$, where p is the ratio of the observed frequency of the helix class over the sample size n. In general, a very good agreement exists between actual and model standard deviations.

We thus calculate the value of δf (the difference in signals) as the sum of all signal perturbations: $\|\delta f\| = \sum_{i \in H} \Delta_i$, where H is the union of the set of helix classes from both the original and perturbed signals. However, while conditioning analysis requires only the size of change $\|\delta f\|$, robustness analysis needs its source. Hence, we also track the total amount of signal change $\|\delta f\|$ as partitioned into three subcategories: signal that stays the signal, signal that becomes part of the larger sample or vice versa, and signal that disappears or appears from the overall universe of helices. These three categories can be interpreted through the lens of robustness: changes that are either signal stable, sample stable, or unstable. These categories, abbreviated as 'signal', 'sample' and 'universe', will become a key part of our analysis to give condition number both an intuitive significance and a threshold for well-conditioned versus ill-conditioned.

4.4 Materials

We now calculate the ratio $\kappa = \sup_{\delta x} \frac{\|\delta f\|}{\|f(x)\|} / \frac{\|\delta x\|}{\|x\|}$ for all ten parameter sets each at 5, 10 or 20% perturbation. Under the supremum requirement of the definition, we set the largest ratio out of the ten parameter sets as the relative condition number κ.

We chose RNA families of differing average lengths (see Table 6), and selected five sequences from each family to span the available range of MFE accuracies. This was done to explore possible correlations between κ with both sequence length and MFE accuracy. Previous results indicate differing behaviors across both sequence length (with respect to prediction accuracy [35]) and MFE accuracy (with respect to SHAPE-directed accuracy [151]); it is feasible that conditioning behavior also be correlated across sequence length and/or MFE accuracy.

Finally, these families were also chosen for their highly structured conformations; their structures are known to be stable under a variety of conditions. Thus, it is presumed that any instability or ill-conditioning of the sampling prediction is due to the algorithm, and not a reflection of the underlying biology.

All Boltzmann samples were generated using GTfold's GTboltzmann function [152].

Name	Length			MFE acc.		
	med	min	max	med	min	max
tRNA	75	73	77	0.51	0.00	0.95
5S rRNA	120	119	122	0.55	0.15	0.85
RNaseP	327	205	354	0.49	0.13	0.68
Intron group I	543	480	554	0.30	0.06	0.74
16S rRNA (small)	958	940	969	0.25	0.14	0.45
16S rRNA (med)	1259	1231	1399	0.29	0.17	0.37
16S rRNA (long)	1537	1528	1548	0.41	0.18	0.64
16S rRNA (extra)	1962	1841	2090	0.34	0.18	0.42

Table 6: Table of RNA families tested, which were chosen to span a range of lengths. The data on tRNA, 5S rRNA and 16S rRNA families were taken from the Comparative RNA Website [21], the data on RNaseP from the RNase P Database [15], and data on intron group I from Rfam [56]. Each family is represented by five sequences that span the available spectrum of MFE accuracies, as calculated by F-measure. The 16S rRNA sequences were subdivided based on length into four categories roughly 300-400 nucleotides apart, as this is the spacing for the two prior families: sequences in the 'small' category are around 950 nucleotides long, those in the 'medium' category around 1250, those in the 'long' category around 1550 and those in the 'extra' long category around 1950. This table provides the median, minimum and maximum lengths and MFE accuracies of the five sequences in each family. Further sequence information can be found at the end of the paper.

4.5 Results

We computed the relative condition number κ for each of the sequences in the families in Table 6. Median condition numbers for each family are given in Figure 2, with subsequent analysis with respect to robustness in Figures 3 and 4. We further investigated the relation of κ to MFE accuracy, length, perturbation level, and signal behavior by means of correlation analysis, demonstrating that κ has a strong and clear correlation to signal behavior. Because signal behavior was explicitly defined in terms of robustness, results thus demonstrate the equivalence of the quantitative condition number and the qualitative measure of robustness, leading to a characterization of sequences that is both rigorous and intuitive.

A number of observations can be made about Figure 20. First, the size of changes in the Boltzmann sample signal is not linear in the degree of perturbation, as the condition number does not remain the same across perturbation levels for any family. Additionally, there is no clear pattern for κ across perturbation levels; while many families see an increase in κ from 5% to 10% to 20% perturbations, Intron group I, 16S rRNA medium and 16S

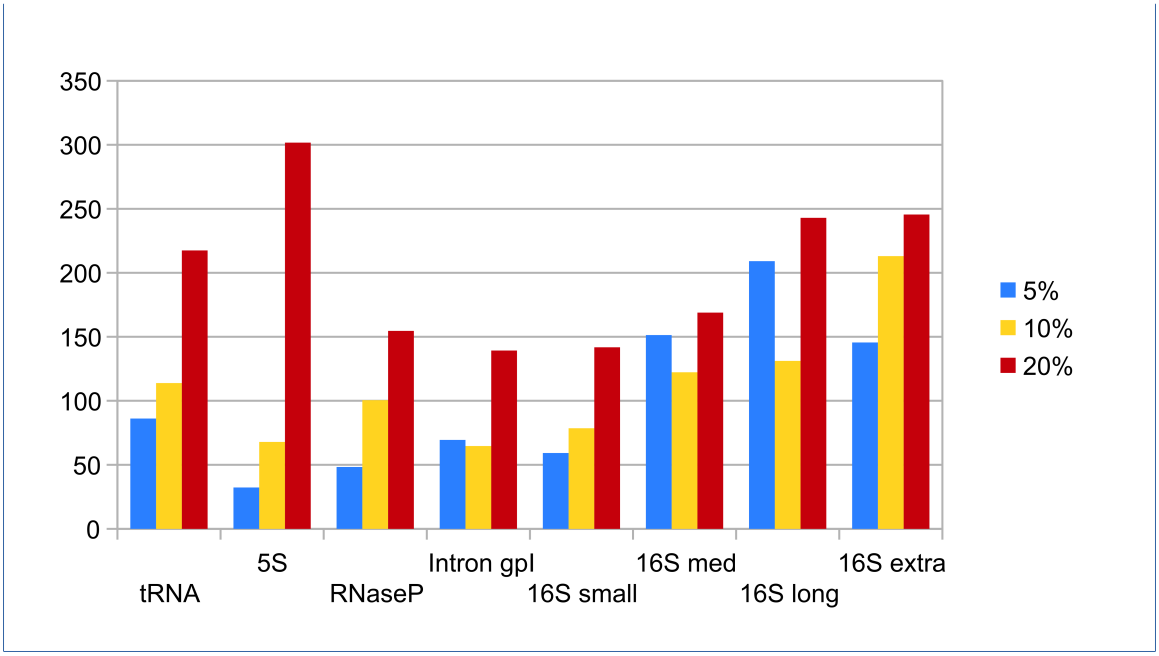


Figure 20: Median condition number for the five sequences in each RNA family. Results are by RNA family and per perturbation level, with RNA families ordered by ascending median sequence length. Similar to prediction accuracy, it is not clear what characteristics of the sequence gives rise to differing values of conditioning.

rRNA long are notable exceptions. Neither is there an obvious pattern at first glance to κ with respect to families of longer or shorter lengths. However, a more in depth analysis confirms that a positive correlation exists between length and κ for both 5% (Spearman's r = 0.4715, p = 0.0021), 10% (r = 0.3313, p = 0.0368), but not for 20% (r = 0.1305, p =0.4222), indicating that for lower perturbations, shorter sequences are better conditioned.

Correlation analysis was also done on κ against MFE accuracies. Although it is not clear why some sequences are either poorly predicted or ill-conditioned, a correlation between them would have had significant implications, since the condition number could then give a confidence estimate of prediction accuracy for sequences for which there are no known structures. Unfortunately, after calculating Spearman's coefficients for all 120 sequences, no signficant correlation was found for any perturbation level, at either 5% (r = -0.1526, p = 0.3471), 10% (r = -0.1077, p = 0.5083) or 20% (r = 0.2395, p = 0.1366). Indeed, we noted the existence of inaccurate sequences with both low and high κ; this fact will be discussed in more depth later. Thus, there is no evidence that the unknown sequence characteristics causing either inaccurate predictions or ill-conditioning are related.

Instead, we found that small κ is related to the robustness of the signal, as partitioned into three categories: that which remains the signal (signal robustness), that which becomes the part of the larger sample or vice versa (sample robustness), and that which either appears or disappears from the sample to the universe of structures (non-robustness). To illustrate this relationship in Figure 21, we take Figure 20 and partition each condition number into these three categories.

Figure 21 shows that the proportion of these three categories differs drastically across sequences. The 'signal' category is a much larger proportion of the total for smaller sequences at lower perturbations; these are also the sequences with lower condition number. At stronger perturbations, the second 'sample' category begins to dominate. Finally, the most unstable 'universe' category is largely not seen until the strongest, 20% perturbation for the longer sequences. These are also the sequences with the largest condition number.

These trends are confirmed when we apply this same analysis to all sequences in Figure 22, and not just the medians of each family in Figure 21. Smaller condition numbers clearly

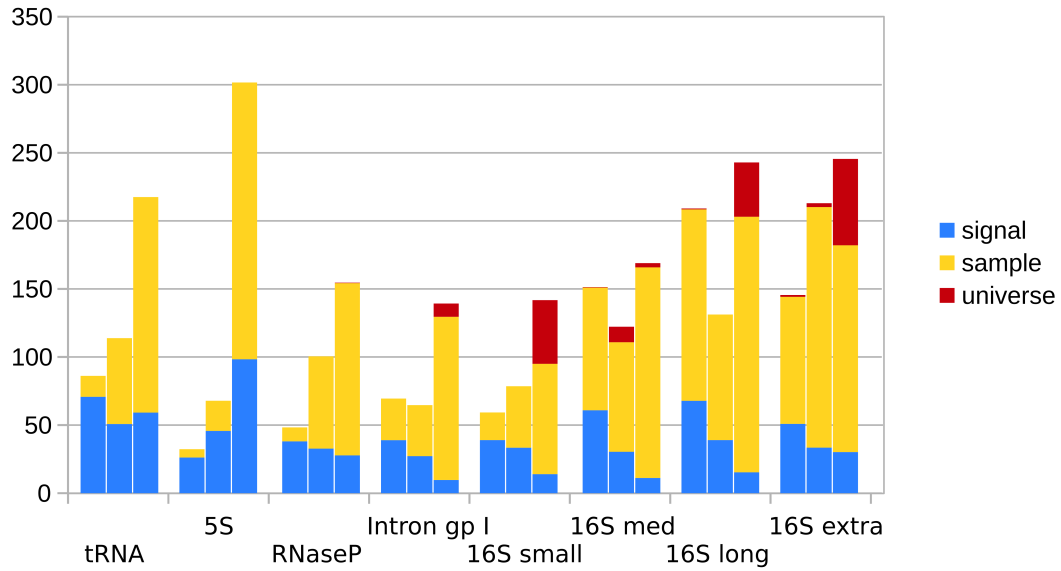


Figure 21: The same values from Figure 20, but subdivided by three categories of changes: those involving movement within the signal ('signal'), those involving movement outside the signal but within the sample ('sample'), and those involving movement outside of the sample within the universe of helix classes ('universe'). Note the dominance of the 'signal' category in sequences of smaller κ, while the 'universe' category only appears in the longer sequences and/or at higher perturbations.

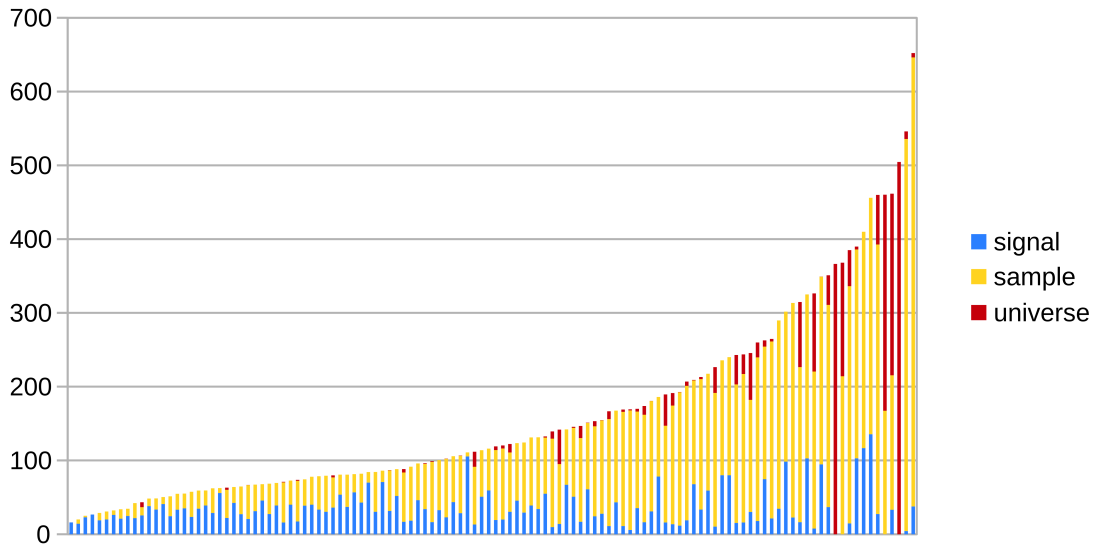


Figure 22: All sequences ordered by ascending condition number. Each condition number is again subdivided into the three categories of Figure 21. The well-conditioned sequences, with a large proportion of blue 'signal' changes, have values less than 90; the ill-conditioned sequences begin at 130, where the red 'universe' changes begin to be more prominent.

have a much larger proportion of blue 'signal' changes. As κ grows, almost all of the growth comes from yellow 'sample' changes; the absolute amount of 'signal' changes stays relatively constant. Changes in the last red 'universe' category begin to appear in significant quantity in the higher values of κ. Thus, Figure 22 indicates strongly that 'signal' changes are associated with low κ, 'sample' changes with moderate to high κ, and 'universe' with high κ.

Correlation analysis quantifies this relation when we compare all forty sequences' κ versus the proportion of each category at three different perturbation levels. We find them to be highly correlated, i.e. the size of κ is predictive of its underlying sources of change. Strong correlations exist between κ and the percentage of 'signal' changes ($r = -0.8082$, $p = 6.6072 \cdot 10^{-29}$), the percentage of 'sample' changes ($r = 0.6149$, $p = 4.3417 \cdot 10^{-14}$), and the percentage of 'universe' changes ($r = 0.5553$, $p = 4.6224 \cdot 10^{-11}$). We shall see that this strong correlation to signal behavior provides an elegant way to interpret κ in terms of robustness, which in turn will aid in defining rough guidelines for well- versus ill-conditioning.

4.6 Discussion

The tight correlation between the mathematical definition of conditioning and the biologically inspired definition of robustness has a number of important implications. Namely, it indicates that the three categories of robustness may also be used to set conditioning thresholds between well-conditioned, ill-conditioned and intermediate sequences. Based on these thresholds, we determine that the majority of these sequences are not ill-conditioned, but instead are sample robust against perturbations. This provides an explicit verification to the long-held implicit belief that Boltzmann sampling mitigates the ill-conditioning of MFE prediction methods. Finally, the existence of both well- and ill-conditioned sequences, coupled with the lack of any correlation with MFE accuracy, implies that both NNTM parameter refinement and also alternate prediction methods should be pursued in order to improve prediction accuracy. The former implication follows from the existence of ill-conditioned, inaccurate sequences, while the latter follows from the existence of well-conditioned, inaccurate sequences.

Because there is a strong correlation between κ and robustness, we use the different categories of robustness—changes that either remain in the signal, remain in the larger sample, or are not confined to the sample— to define the different categories of conditioning. Namely, we use the observation for Figure 22 that signal robust changes in blue dominate for early values of κ, sample robust changes in yellow in the midrange of κ, and changes not restricted to the sample in red for higher values of κ.

Because such a strong relation exists, we use the different robustness categories to define specific thresholds for well- versus ill-conditioning. Intuitively, well-conditioned sequences should correspond to sequences in which the majority of changes occur within the signal. To find the range of such sequences, we calculate the average percentage of 'signal' changes over a window of five consecutive sequences; we set the well-conditioned threshold to the last value in which the average for the preceding five values is above 50%. This turns out to be at the 48th sequences, which has a κ of 88.182.

Similarly, in order to find the threshold for ill-conditioned sequences, we calculate the average percentage of the most disruptive 'universe' changes for a sliding window of five

sequences. We set the ill-conditioned threshold at the point at which the average goes above 10% for the first time; this is at the 70th sequences with a κ of 131.257.

Thus, sequences with κ less than 90 can be considered well-conditioned, with a signal that will likely remain the signal even under perturbations. Similarly, "semi-conditioned" or intermediate sequences with κ between 90 and 130 are likely to be sample stable; i.e. while the entire signal is not likely to remain signal under perturbation, the overall sample is merely experiencing a reweighting of its frequencies. Finally, sequences with condition numbers above 130 should be considered ill-conditioned; it is likely that a significant part of their changes come from completely new helix classes appearing in the new signal. Thus, the qualitative definitions of robustness married to the quantitative rigor of conditioning provide a clear and balanced analysis of Boltzmann sampling under NNTM perturbation.

The ill-conditioned threshold occurs at the 70th sequence out of 120. That more than half of the sequences are at least sample robust has at least two major implications: first, that the use of Boltzmann sampling against parameter fluctuations is validated, and second, that efforts to refine NNTM parameters in hopes of improving accuracy may be of limited effectiveness.

The first implications follows from the fact that the majority of the sequences merely experience a reweighting of helices under perturbation. Indeed, even much of the ill-conditioned minority have large proportions of sample stable changes, despite some unstable changes. Only 17 of the 120 sequences experienced disruptive 'universe' changes contributing more than 10% of the total; more than 85% of sequences had at least 90% of changes resulting from helix classes already in the sample shifting frequencies, i.e. sample robust helix classes. Thus, while predicting the MFE structure may be considered ill-conditioned [94], sampling from the Boltzmann distribution is arguably more well-conditioned than not, as has long been implicitly assumed but not verified.

The overall sample robustness also has a second implication for accuracy and ongoing efforts to improve prediction methods: both NNTM model improvement and other alternative methods are necessary. Because there was no correlation of κ with MFE accuracy, we know that well-conditioned sequences are not necessarily accurate; they can be stable

around inaccurate low energy structures. Indeed, for the sequences in the well-conditioned, robust category, the median MFE accuracy is 0.34 out of 1; more than a fifth of the well-conditioned sequences have an MFE accuracy of less than 0.2.

Hence, for well-conditioned but inaccurate sequences, minor adjustments to the NNTM may not substantially change the inaccurate predictions; this extends previous results, which have indicated that refined parameters do not uniformly increase prediction accuracies of sequences [35].

Hence, the precision of NNTM parameters is not the only factor affecting secondary structure prediction accuracy; other factors like kinetic traps [72, 119, 157] and multiple native conformations [103, 113, 28, 90] still necessitate the development of alternate and/or complementary computational and experimental methods [34, 172, 2, 107, 89].

However, the existence of ill-conditioned sequences, comprising a third of all sequences, also indicate that efforts to improve the thermodynamic model do remain important. For these sequences, perturbations result in a significant number of new helix classes; some amount of parameter adjustments or improvements will result in a substantially different signal. For sequences with a low MFE accuracy, this may be the difference between an accurate versus inaccurate prediction. Thus, efforts to refine the NNTM are still important, especially when considering longer sequences at higher perturbations, as almost all of these ill-conditioned sequences are.

It is worth mentioning that some exploratory work was done in conjuction with this study in which we perturbed only subsets of the parameters. Results indicate that the majority of the changes tracked by κ came from perturbing either the loop or the stack parameter files; perturbing the other parameters had only a minimal effect. Hence, refining these parameters are likely to pay the biggest dividends in efforts to improve the NNTM. This line of questioning is paralleled and expanded in recent work [176].

Preliminary studies [176] have also indicated that the majority of the tabulated error ranges for the loop and stack parameters fall within the 20% perturbation levels of this study. Thus, the level of perturbations reasonably expected to exist in the loop and stack parameters have been shown here to have a significant effect on a number of sequences.

4.7 Conclusion

For the first time, conditioning for Boltzmann samples is rigorously quantified with a relative condition number κ, and is shown to be highly correlated with robustness. Using this correlation, we define well-conditioned sequences as those that are signal robust with κ below 90, ill-conditioned sequences as those that are not robust with κ above 130 , and intermediate sequences as those that are sample robust with κ inbetween.

Of particular interest are the entirely new helix classes under perturbation that tip sequences into ill-conditioning and non-robustness. They have at least two implications. First, because they make up only a small fraction of all perturbed signals, we conclude that Boltzmann sampling as a whole is robust against NNTM perturbations, in vindication of one of its original purposes. Secondly, because they do exist, this implies that ongoing efforts to refine the NNTM still matter to certain sequences. The lack of correlation between κ and MFE accuracy, however, also indicates that for some well-conditioned but inaccurate sequences, other methods besides NNTM refinement (such as multiple sequence analysis [127, 60, 3] chemical footprinting [167, 47] or SHAPE analysis [27, 110, 146]) need to be pursued to increase accuracy.

As the first study to tackle the conditioning and robustness of a Boltzmann sample for perturbations across the model, this work naturally opens the door for further research. Avenues to be explored include using more sophisticated perturbation models, such as those reflecting parameter dependencies, as well as testing the correlation between sample conditioning and responsiveness to experimental or biological data like SHAPE [151]. Relationships between conditioning and the accuracies of entire samples also remain an open question. With the foundational concepts and metrics introduced in this paper, deeper research into these important yet poorly understood areas has now become possible.

4.7.1 Author Contributions

E.R., D.M. and C.E.H participated in early data discovery and discussions concerning NNTM perturbations and the use of condition numbers. D.M. ran tests on helix class frequency models, and produced Figure 19. E.R. selected the sequences, ran the later data,

and produced Figures 20, 21 and 22, with input from C.E.H. Manuscript written by E.R. and C.E.H.

4.7.2 Acknowledgements

We thank A. Abhishek for his code to generate perturbed NNTM parameters. We also thank D. Mathews for providing both data on observed NNTM errors and comments on the manuscript.

4.7.3 Conflict of interest

None declared.

4.7.4 Funding

Funding for this paper was provided in part by the Burrows Wellcome Trust [CASI #1005094 to C.E.H.].

Family	Name	Accession no.	length	MFE acc.
tRNA	*S. meliloti*	AL591786	77	0
	P. aphrodite, formosana	AY916449	73	0.954
	C. diphtheriae	BX248359	73	0.755
	B. cepacia	L28151	76	0.205
	S. cerevisiae	J01381	75	0.51
5S rRNA	*M. fossilis*	V00647	120	0.15
	M. glyptostroboides	M10432	120	0.29
	S. pombe	K00570	119	0.85
	O. sativa	M18170	119	0.55
	P. waltl	X16851	122	0.76
RNaseP	*T. syrichta*	L08801	286	0.13
	Z. bailii	AF186231	205	0.68
	A.ferrooxidans	X16580	327	0.59
	P. fluorescens	M19024	354	0.49
	H. chlorum	U64881	342	0.32
Intron group I	*S. anglica*	Z69912	554	0.06
	H. rubra	L19345	543	0.30
	T, thermophila	V01416	506	0.74
	P. thunbergii	D17510	550	0.13
	B. yamatoana	D38239	480	0.51
16S rRNA (small)	*S. aestuans*	AJ012746	968	0.34
	A. cahirinus	X84387	940	0.20
	L. catta	AF038013	954	0.251
	N. robinsoni	U93061	969	0.447
	V. ursinus	U61078	958	0.135
16S rRNA (med)	*V. acridophagus*	AF024658	1399	0.371
	V. corneae	L39112	1259	0.33
	E, cuniculi	X98467	1295	0.17
	V. imperfecta	AJ131646	1231	0.288
	E. schubergi	L39109	1252	0.23
16S rRNA (long)	*E. coli*	J01695	1542	0.41
	S. griseus	X61478	1528	0.322
	M. hyopneumoniae	Y00149	1537	0.639
	M. leprae	X56657	1548	0.179
	C. testosteroni	M11224	1536	0.524
16S rRNA (extra)	*O. cuniculus*	X06778	1863	0.177
	R. carriebowenis	AF006089	1841	0.338
	P. falciparum	M19172	2090	0.423
	Z. mays	X00794	1962	0.258
	P. vivax	U07367	2063	0.385

Table 7: Table of RNA sequences tested by family. Note the range of both sequence lengths and MFE accuracies.

CHAPTER V

PREDICTING RNA CONSENSUS STEMS THROUGH UNSUPERVISED CLUSTERING OF UNALIGNED SEQUENCES

5.1 Abstract

Motivation: Improvements in secondary structure prediction accuracy for a single RNA sequence, notably through Boltzmann sampling, have not been realized for multiple homologous ones; consensus structure prediction remains a significant challenge in computational biology. To close this gap, two insights are critical. One is finding the right balance between improvements in precision versus recall. Another is resolving conflicting base pairing signals through an appropriate level of structural granularity. Together, these can achieve very high accuracy predictions of native structural elements for an RNA family.

Results: `ConsensusStems` leverages RNA profiling and noise-sensitive clustering to extract common base pairing regions from Boltzmann samples for related sequences. This focus on more general structural element prediction is very successful; the cluster centroids output recover the native stems in 7 of 11 Rfam families tested, with median sequence lengths up to 300 nucleotides and structural complexity up to 10 stems. Overall, the (avg, std) centroid accuracies are: precision = (0.96, 0.08), recall = (0.95, 0.10), and approximate Mathews correlation coefficient = (0.95, 0.08). Thus, `ConsensusStems` is a important contribution to advancing RNA folding prediction.

Availability: A demonstration webserver is online at `rnaconsensus.math.gatech.edu`. Code can be downloaded for general use via `https://github.com/gtfold/ConsensusStems`.

Contact: heitsch@math.gatech.edu

5.2 Introduction

Accurate prediction of the common native structure for homologous RNA sequences is an open problem in computational biology [46, 45, 170, 60, 3, 92]. Given current interest in

'noncoding' RNA's role in gene splicing, editing, and regulation, this challenge has taken on new urgency in recent years. In particular, since experimental determination of 3D conformations is still time-consuming and labor-intensive, function is most often inferred from computational predictions of RNA secondary structures. Thus, improved prediction of the noncrossing, canonical base pairings common to related RNA sequences is essential to providing new functional insights.

Although sequence alignment is a typical starting point for consensus structure prediction, such methods face the difficulty that RNA pairings (i.e. complementarity of two positions i and j) are much more strongly conserved than individual nucleotide identity. In contrast, aligning RNA secondary structures rather than primary sequences can produce a more complete and accurate consensus prediction.

It is not obvious, though, which structural elements to align; minimum free energy (MFE) structures [62, 102], base pair probabilities [63, 145], sampled helices [173], or even all stable helices [76, 59, 153, 4, 57] have been tried. The key is striking the right balance; too little information, and the recall limitations of the original false negative predictions cannot be overcome. Too much, and the resulting precision is dominated by false positive predictions. Our novel `ConsensusStems` approach achieves a good balance by leveraging the predictive power of Boltzmann sampling [32] filtered through the denoising achieved by RNA profiling [134].

Stochastic sampling from the Boltzmann ensemble for a single RNA sequence is state-of-the-art in secondary structure prediction since it efficiently provides the most comprehensive folding information [105, 142]. Hence, by starting the `ConsensusStems` pipeline with Boltzmann sampling, false negatives are minimized to the maximum extent possible under the current nearest neighbor thermodynamic model (NNTM).

False positives are then filtered by RNA profiling which extracts the structural 'signal,' i.e. the set of high frequency maximal helices known as features, from each noisy sample. This set of features is a robust signature of each Boltzmann ensemble that is easily compared between related sequences to highlight similarities and differences [133].

Such comparisons are the core of `ConsensusStems`' methodology since the NNTM prediction for each sequence is only partially correct on average [132]. However, these partial signals almost always include complementary information. Thus, false negatives are minimized by consolidating the individual sets of features over the entire family. This recovers the native helices, albeit with considerable noise.

To improve precision as well as recall, the false positives are filtered by a noise-sensitive clustering algorithm [43]. The resulting clusters are the structural 'alignment' produced by `ConsensusStems`. Specifically, each cluster is denoted by a representative centroid; this is the consensus stem prediction that those 5' and 3' segments interact exclusively with each other. Each output cluster also has an associated list of supporting sequence/feature pairs, consolidated into a sequence-specific stem. These are the regions of the individual sequences understood to be structurally 'aligned' with the consensus stem.

This focus on predicting regions of interaction, i.e. the forest rather than the trees, is critical to `ConsensusStems`' success. NNTM optimization often predicts very similar helices, presenting a conflicting base pairing signal that is challenging to resolve accurately. However, at a lower level of structural granularity, competitors tranform into allies, sending a clear true positive signal for a native pairing region — which could well a better reflection of physical reality given the stochasticity of biological systems. As will be shown, our stem abstraction clarifies base pairing patterns and increases prediction accuracy of consensus structure for multiple sequences.

By minimizing false positives as well as negatives, `ConsensusStems` achieves remarkable accuracy at this level of granularity. Tests on a diverse set of 11 different RNA families demonstrate that this new method predicts the native consensus stems with an average accuracy over 95%. It is 100% accurate 66% of the time, and the remaining four families were 89%, 89%, 86%, and 75% accurate. Hence, this approach is a major step forward in resolving the consensus structure prediction problem.

5.3 Approach

Our goal is predicting common regions of structural interation, called consensus stems. Results will show that while the signal at the base pair level is usually messy, viewing the same data at a lower granularity yields clear and accurate predictions.

More precisely, we generalize the standard (i, j, k) notation for helices, which denotes k consecutive base pairs closed by (i, j), to the stem notation (i, j, k, l). Stems have an extra coordinate l, since the length of the 5' region (k) may not be equal to the length of the 3' region (l). The stem coordinates (i, j, k, l) thus denote that the regions $[i, \ldots, i + k - 1]$ and $[j - l + 1, \ldots, j]$ interact exclusively with high probability.

Accurate prediction of native consensus stems requires dealing with false positives (or FP, the non-native predictions) and false negatives (or FN, the native elements not predicted). `ConsensusStems` does this in two rounds: (1) by using Boltzmann sampling to deal with FN, and profiling to deal with FP; and (2) by using multiple, normalized sequences in the same family to deal with FN, and clustering to denoise the composite family data to deal with FP.

Proof-of-principle for this approach is given in this section by analyzing a set of tRNA sequences. Although the native cloverleaf is well-known, minimum free energy (MFE) prediction accuracies for an individual sequence can range from a high of 100% to a low of 0% [134]. Hence, a consensus method which starts with a single MFE structure per sequence is likely to have a prohibitive number of FN. Hence, current best practice is to stochastically sample a set of predicted structures (typically of size 1000) from the Boltzmann ensemble for that sequence [32].

The Boltzmann sample, however, contains many more structures than the native, and necessitates post-processing to deal with FP. RNA profiling has been demonstrated to extract the key structural information from a sample [134], yielding a clear, concise — although not necessarily fully correct — signal from the noisy ensemble. High level patterns are readily seen because of the key use of abstraction, which improve computational accuracies significantly [132].

Helix	Triplet	Frequency
1	1 71 7	1000
2	24 44 7	834
3	21 47 2	760
4	46 66 7	544
5	15 64 4	340
6	50 60 4	193

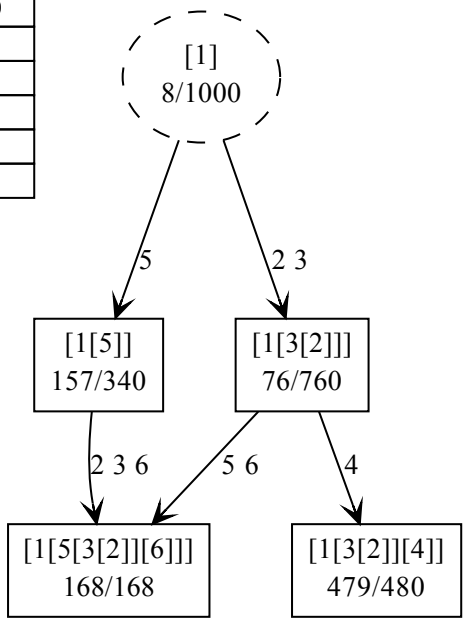


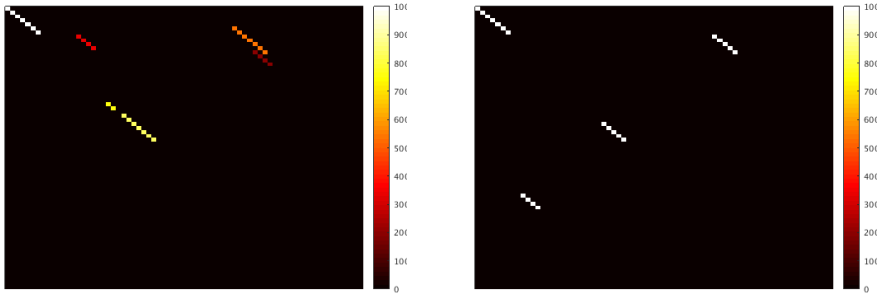
Figure 23: Profiling output for *T.brucei*. Maximal helices are listed in descending frequency with (i, j, k) triplet and corresponding index number. Profiling uses a maximum average entropy threshold to truncate the distribution, returning only the most common helices as the selected 'features.' Each node in the graph gives a profile, i.e. a maximal combination of features, with brackets indicating nesting relationships. The ratio gives the number of sampled structures with exactly that profile (numerator) over the number with at least those features. Nodes are related as a Hasse diagram under the partial ordering of set inclusion, with edges labeled by the difference. For this sequence, the most frequent profile is [1[3[2]][4] which was sampled 479 out of 1000 times and is nearly the native structure. The FP is feature #3 at $(21, 42, 2)$ with estimated probability of 76.0%. The FN of $(10, 24, 4)$ is the 11th most frequent helix with a probability of only 5.9%.

5.3.1 Profiling *Trypanosoma brucei* lysine tRNA

We use a representative tRNA sequence *Trypanosoma brucei* (Accession Z11880.1/124-195) from the Rfam seed alignment [56] to see that profiling successfully extracts enough native signal from a sample to be our starting point.

Figure 23 shows the list of features, i.e. high frequency maximal helices, and summary profile graph of *T.brucei*. By consolidating substructures with high similarity and truncating the long, noisy tail of the frequency distribution, profiling produces a clear, concise, and stable structural signal for this ensemble.

As seen in Figure 24, this signal contains a significant portion of the native cloverleaf;



(a) Z11880.1/124-195 (b) Native

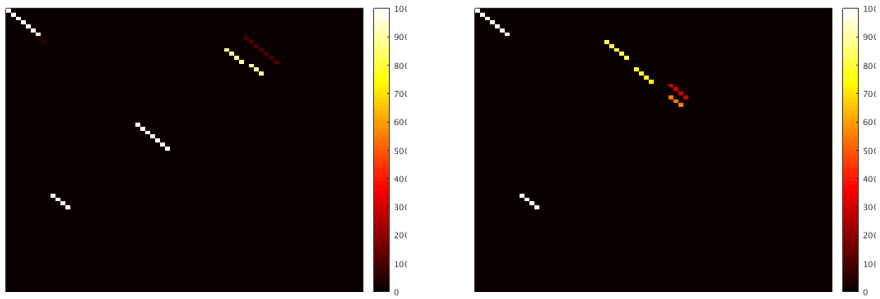
Figure 24: Two dotplots for *T.brucei*: the 6 features extracted by profiling from the Boltzmann sample (left) and the native secondary structure (right). A base pair between positions i and j corresponds to a box with coordinates (i, j) in the (x, y) plane, with $i < j$. On the left, the colors correspond to a frequency heatmap from red/least to white/most. It is clear that the native structural signal is partially present in this ensemble, albeit noisy and incomplete.

3 of 4 TP helices are high frequency substructures in the *T.brucei* ensemble. However, without additional information, it is not possible to distinguish the true from false positives among the 6 features output by profiling. Likewise, although the FN helix is present in the whole sample with almost 6% frequency, there is no reason to identify this particular helix from the 44 others that are truncated from the full distribution.

This demonstrates that, although the complete native structure is seldom present with high probability in a single Boltzmann ensemble, there typically exists a significant amount of *partial* information. Moreover, as shown below, different sequences capture different parts of the native structure among their features. Hence, a consensus structure can be recovered by agglomerating the helix signal from homologous samples to produce a composite signal with (very) high accuracy.

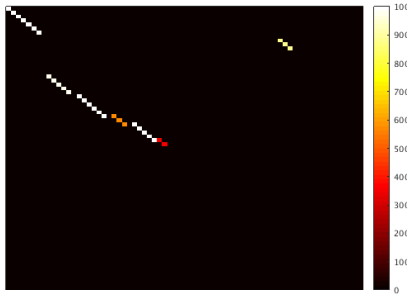
5.3.2 Leveraging information from homologous sequences

We now illustrate that (1) profiling extracts part of the native structure from a sequence reasonably consistently; and (2) the partial signals from a large enough set of homologous sequences are complementary. Hence, the common native structure can be recovered by amalgamating individual sequences' features.

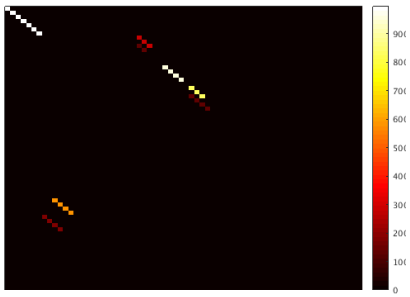


(a) U18089.1/221-293

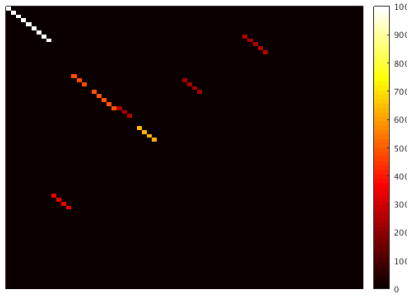
(b) AF008220.1/6620-6692



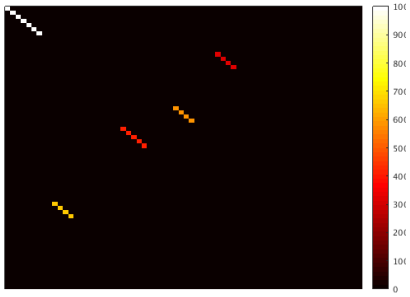
(c) X55321.1/1-72



(d) D85268.1/459-527



(e) AB042809.1/5159-5089



(f) D16387.1/11325-11257

Figure 25: Heatmaps for the features of six tRNA sequences. Each square (i, j) corresponds to the base pair (i, j), with the frequency of the base pair (as measured by frequency of the maximal helix to which it belongs) reflected in the color, from the highest frequency (white) to the lowest (red). While not all the sequences have the native cloverleaf structure in the features (see Figures 24), all have at least some native helices as a feature.

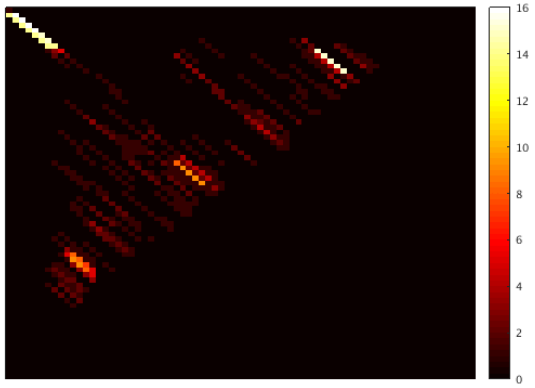


Figure 26: Representation of the 2-D normalized grid for tRNA to which all high frequency helices are mapped. Each helix (i, j, k) is mapped to its corresponding grid points, augmenting frequency counts for cells $(i, j), \ldots, (i + k - 1, j - k + 1)$. The frequency of each cell is represented by color, with black indicating zero counts, through red up to white, the highest count. Note the general shape of the tRNA cloverleaf structure, a closing stem encompassing three stacks (see Figure 24), is present though somewhat blurry.

Figure 25 shows six additional tRNA sequences with a typical range of features. Although it is possible that the native structure is fully predicted, as for *U18089.1*, most have only a partial signal. However, the collective signal is sufficient to recover the consensus stems, as seen in Figure 26.

Figure 26 is a composite heatmap of 30 randomly chosen tRNA sequences (including those from Figure 25). The features for each sequence are are overlayed on top of each other after normalizing for differing lengths. Figure 26 clearly contains the complete tRNA structure.

The consensus helices of Figure 24 are clearly seen in Figure 26, albeit with a signficant amount of noise. Thus, recovering the native tRNA stems at this point is matter of filtering out the FP 'noise'; this insight is the governing principle of `ConsensusStems` in using clustering to automate the extraction of the native signal. This correspondence between the consensus stems and the strong composite ones is found in general for RNA families, as the next section shows.

5.3.3 Examining Rfam families

We demonstrate that tRNA is not unique by considering eleven families from Rfam [56] that span the range of sequence lengths. While many of these sequences have been used as test sequences in the literature (Table 8), no standard test set has been developed to benchmark consensus structure methods. These eleven families were selected to span the range of sequence lengths available from the set of Rfam families with known structure. Sequences from the families were randomly chosen from the seed alignment, sampled and then profiled to obtain their features.

Family	Lengths		Sequence		Structure	
	Med	Range	Ident	Num	Helices	Stems
tRNA+	72.5	22	46	30	4	4
THF	98	21	62	25	5	3
TPP*+	105	89	56	29	6	5
5S*+	117	18	60	29	11	3
FMN	138	76	72	28	5	5
U1	162	18	65	25	12	5
ykoK+	168	25	61	26	13	5
glmS+	173.5	70	60	18	6	4
IRES cripavirus	199.5	36	53	24	8	4
IRES HCV*	243.5	185	86	24	20	10
metazoa SRP	298	27	70	23	18	7

Table 8: Information for 11 test families, including average length, average family pairwise sequence identity in percent, number of seed sequences analyzed, and number of helices and stems in Rfam's secondary structure. Sequences from each family were randomly chosen, and each family was chosen to span the range of lengths available from the set of Rfam families with structures. An asterisk indicates families included in the MASTR data set [101], a popular benchmark limited to shorter sequences. A plus sign indicates a family used by RNAscf [4], a method also working with helices.

Figure 27 demonstrates that while a percentage of native helices in each family are low frequency to non-existent, the majority are strongly present in the sample as features. This continues our findings that a significant though partial native signal exists within the features recovered by profiling for each sequence.

Furthermore, we show that the composite signal of multiple Boltzmann samples recovers the native structure with noise for all families tested. Hence, our strategy can be summed in two parts: consolidate the signal by agglomerating individual Boltzmann samples to address FNs, and filter the signal through clustering to address FPs.

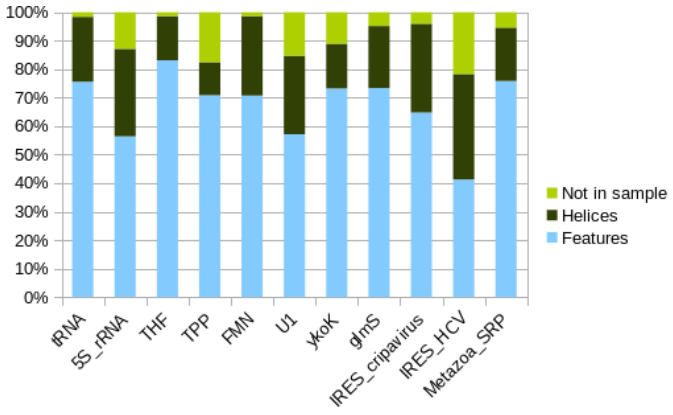


Figure 27: Data for 11 Rfam families, indicating within a family the number of native helices with multiplicity for which it is a high frequency feature (light blue), a low frequency helix (dark green), or not present in the sample at all (light green). For k sequences in a family whose native structure has n native helices, the total number of native helices categorized is nk. For most of the families, the majority of native helices are high frequency features. Only a fraction are not present in the sample at all.

5.4 Methods

Figure 28 illustrates the general steps of `ConsensusStems`, which starts with Boltzmann samples for a set of related sequences, and ends with a list of clusters composed of sequence/feature(s) pairs. Each cluster is characterized by a centroid stem, the generalized (i, j, k, l) coordinates defining pairing regions, which is the final prediction by `ConsensusStems`.

1. *Generate* a Boltzmann sample for each sequence in the family.
2. *Profile* each sample to get the sequence specific features.
3. *Cluster* all the feature to get the potential consensus stems.
4. *Refine* each cluster by adding in features from missing sequences.
5. *Validate* each cluster by assessing overall base pairing support for the region across sequences.
6. (a) If there are new clusters found, *resample* the structures with a constraints file generated from the clusters; go back to step 2.

   (b) If there are not any new clusters found, *terminate* the procedure.

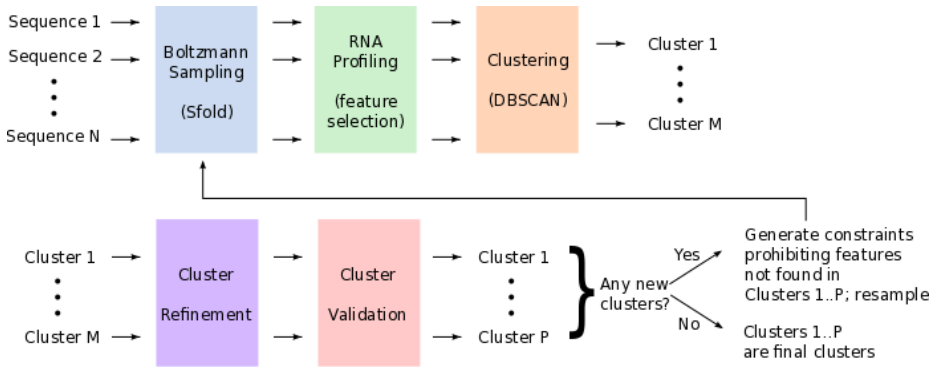


Figure 28: The RNA `ConsensusStems` method. Each sequence in a family is sampled and profiled, yielding a set of features that are then normalized and clustered. The initial clusters are then refined by searching for potential additions from missing sequences. Finally, they are validated by assessing each sequence's possible base pairings in the region of interest. If any new clusters are identified, then the final clusters are used to make a constraints file that feeds back into Boltzmann sampling.

5.4.1 Step one: generate a Boltzmann sample

`Sfold` 2.2 was used with default settings to generate a standard Boltzmann sample of a thousand structures for each sequence in a family. Although various programs exist that implement Boltzman sampling, `Sfold` was used because of its option to sample with constraints, which option will be used later in `ConsensusStems`.

5.4.2 Step two: profile the samples

The output of RNA profiling gives a list of all the features with its (i, j, k) coordinates (see Figure 23), which denote a set of consecutive base pairs $(i, j), (i+1, j-1), \ldots, (i+k-1, j-k+1)$.

5.4.3 Step three: cluster the features

While many clustering methods exist [80], one is necessary to filter out potential 'noise' in order to recover the native signal (see Figure 26).

We chose to use the clustering method `DBSCAN` (Density-based Spatial Clustering of Applications with Noise) with its inherent concept of noise, one of the most commonly used and cited clustering algorithms [43, 78]. This algorithm classifies data points as a *core point*, *reachable point*, or as *noise* (Figure 29), using only two parameters: a radius ϵ and a

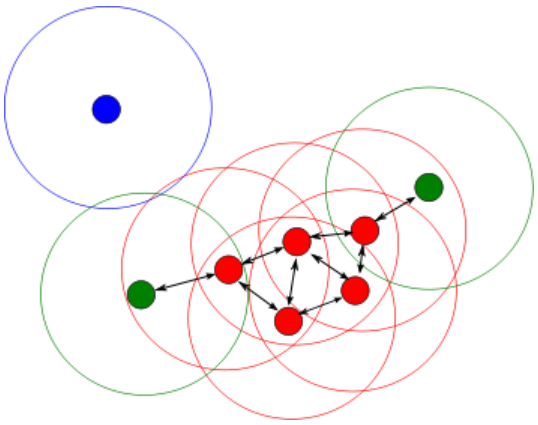


Figure 29: Schematic of the clustering method DBSCAN. The radius ϵ is denoted by the circles, whose colors correspond to the point on which it is centered. Each red point has $P = 4$ points within its radius (including itself) and is a core point. All points in a circle's radius are *reachable* from the core point, and belong to the same cluster as the core point. Each of the green points are reachable from a red core point, and hence are part of its cluster, but are not themselves core points. The blue point is neither a core point nor reachable from a core point; it is considered noise and not part of a cluster.

minimum number of points P.

- A *core point* has P points (including itself) within a radius of ϵ, and is part of a cluster.
- A *reachable point* is within ϵ of a core point, and is part of the same cluster as the core point.
- A point is *noise* if it is neither a core point nor a reachable point, and is not assigned a cluster.

The steps for denoising the signal and producing a clustering of all the features is thus:

1. Cluster the 1D distribution of sequence lengths to find the radius ϵ
2. Construct normalized 2D space of all features
3. Cluster the 2D distribution of base pairs in all the features
4. If majority of points are noise, adjust P and run step 3 again

5.4.3.1 Cluster sequence lengths to determine ϵ

For families with very similar lengths, homologous helices are located close to each other in the normalized space. Hence, the radius ϵ should be small to avoid including FP in the cluster. Conversely, for families with a wide range of lengths, a more generous radius ϵ should be used to avoid excluding FN from a cluster too narrowly defined.

We run a 1D DBSCAN on the distributions of lengths using initial $P = \frac{N}{4}$, with N being the total number of sequences in a RNA family. We begin with $\epsilon = 0$, to allow for the case that all sequences are of the same length. If the majority of the lengths are then classified as 'noise', we increment ϵ by 1 and run DBSCAN again with the new parameters. We stop as soon as we have found the lowest ϵ which produces a clustering of lengths in which the majority are either core or reachable points.

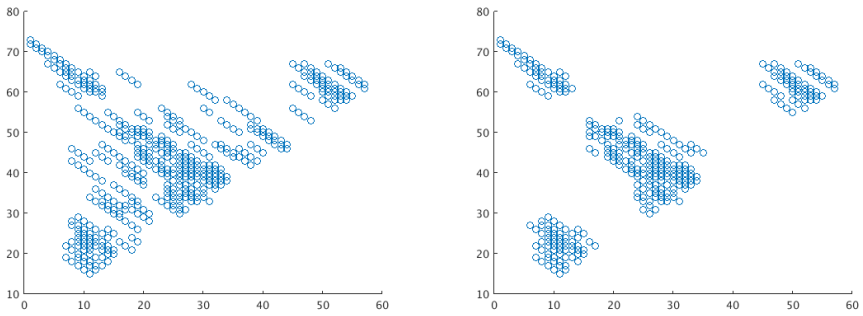
5.4.3.2 Construct the normalized 2D space

Many sequences in the same family have differing lengths, often by a significant amount due to insertions or deletions in the sequence. Hence, all feature coordinates are normalized to the median length in order to embed them onto a common clustering space. Given a sequence S with length d belonging to a family of median length n, the coordinates of the features of S are multiplied by $\frac{n}{d}$ and rounded up.

In an nxn grid, the (i, j) square is associated with the base pair (i, j). Each square (i, j) has an associated frequency indicating the number of features containing the (i, j) base pair; computationally, the (i, j, k) feature helix causes the frequency of squares $(i, j), \ldots, (i + k - 1, j - k + 1)$ to each be augmented. This grid will be the target of the clustering method. The frequency of all features of tRNA is emphasized in Figure 26, while the location is emphasized in Figure 30.

5.4.3.3 Cluster the 2D features

Before running DBSCAN to cluster the points, the distance metric needs to be considered in light of the biology of insertions/deletions (called 'indels'). The distance between two coordinates can be considered the number of indels necessary to shift one into the other.



(a) All tRNA features (b) After initial clustering

Figure 30: Figure on the left is the normalized space of all the features of tRNA. The figure on the right represents all the features found to be in a cluster after initial clustering of step two; unsupported features have been filtered out as noise.

Hence, the Manhattan distance metric is used: a point (i', j') is considered within a radius ϵ of (i, j) if $|i - i'| + |j - j'| \leq \epsilon$.

Since an indel (relative to the median) of d nucleotides can occur in the sequence before i, this will cause an offset of d relative to both i and j, with an overall distance of $2d$ from the original (i, j) point. To allow for this, we set a new $\epsilon' = 2\epsilon$ for the 2-D clustering.

5.4.3.4 Adjust parameters if needed

Clustering is run with parameters ϵ' and P on the 2D grid. If the majority of points on the 2D grid is labeled as 'noise', clustering is run again with ϵ' held constant and P decremented by 1. This cycle repeats until a $P' \leq P$ is found such that the majority of the points are either core or reachable points. At this point, the initial clustering of points is complete (Figure 30), with each cluster being a list of sequence/feature(s) pairs. This list can be considered an implicit alignment, with the features of each sequence in the cluster structurally aligned with each other.

5.4.4 Step four: Cluster Refinement

Each cluster is refined by searching all the missing sequences for potential cluster members. The initial clustering could miss these features due to large differences in sequence lengths that even normalization does not fully address. Hence, in order to refine the cluster and fill

in the gaps, each cluster:

1. Finds all the sequences not present in the cluster
2. Determines coordinates of each missing sequence's search window
3. Identifies all features that fit the parameters of the search window
4. Calculates probabilities of implied indel positions of found features
5. Adds feature to cluster if probability is over a threshold

5.4.4.1 Find missing sequences

Each cluster is associated with a list of sequence/feature pairs. All sequences not represented in the cluster are the missing sequences, which are examined one by one.

5.4.4.2 Determine search window

Each missing sequence is scanned for features that could plausibly be structurally aligned with the others in the cluster. The idea of plausibility is rooted in the observation that the relative length of a sequence can be roughly correlated with the relative displacement of its stem. Namely, a longer sequence tends to have 'later' stem coordinates than those of shorter sequences in the alignment, and vice versa for shorter sequences. This insight helps us to define a window for each sequence in the unnormalized, original coordinates of the sequence, in which the native stem is expected to be located.

More specifically, the location of each sequence's search window is based on the sequence's expected number of indels. This can be calculated from the displacement of the length of the sequence relative to the cluster.

To obtain a reference point for a cluster, its centroid is calculated as the median coordinates of all constituent stems (i, j, k, l) in the cluster. We also calculate the median length d_c of all sequences included in the cluster.

Given a missing sequence S_m of length d_m, any potentially homologous helix of S_m is likely to be located an offset of $\delta d = d_m - d_c$ away from the centroid stem. This window ranges from an offset of $1.25\delta d$ to $-0.25\delta d$, for reasons explained below.

As an example, consider a cluster of median length $d_c = 75$ and centroid $(10, 30, 5, 6)$, and a missing sequence S_m of length $d_m = 80$. S_m can be expected to have at least $\delta d = d_m - d_c = 5$ insertions not present in sequences of length d_c. Since up to $\delta d = 5$ insertions could occur before the centroid, a window of up to δd after the centroid needs to be searched. Actual data demonstrates that a few indels often occur in the opposite direction (i.e. deletions in our example). To allow for this, up to 25% of δd (rounded up) is 'budgeted' in the window in the opposite direction. The window offset needs to be increased from δd to $1.25\delta d$, in order to balance out the $0.25\delta d$ in the opposite direction. For our example, this means that to find potential homologous features to the original centroid pairing region $[10, 14]$ with $[25, 30]$, we would look for those pairing the regions $[10 - 0.25\delta d, 14 + 1.25\delta d]$ with $[25 - 0.25\delta d, 30 + 1.25\delta d]$.

5.4.4.3 Identifying potential features

For every missing sequence, its Boltzmann sample is scanned for features forming pairings between the two regions of the determined window.

5.4.4.4 Calculate probabilities

The search window may identify many FP candidates that still need to be filtered out, especially if the difference in sequence lengths is large. Hence, additional filtering is done based on the candidate feature's implied locations of indels that enable its structural alignment to the cluster.

Features are only accepted into a cluster if the implied number of indels falls within acceptable boundaries of plausibility, e.g. if the ith coordinate is shifted by m indels compared to the centroid, we expect the jth coordinate to be shifted by at least m indels.

More precisely, the occurrances of indels can be modeled as a Poisson process, assuming the rate of indels as independent and identically distributed for simplicity. (We assume that if the centroid coordinate at i_c and its putative homolog in a sequence is at position i_m, then there have been $m = i_c - i_m$ indels. There can be many more, of course, as long as the net displacement equals m, but likelihood of this is far less.) In the spirit of the simple gap penalty, a uniform rate of indels is assumed for simplicity, given by $k = \frac{\delta d}{d_m}$, where δd is

the absolute difference between a sequence's length d_m and the median length of the cluster d_c. Given this rate , the Poisson probability of observing $\lambda = m$ indels over p nucleotides is calculated:

$$P(\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$$

The Poisson probability is then normalized over the largest probability, where $\lambda = pk$, which is the rate of indels k times the length of interest p.

Poisson probabilities are calculated for three intervals: the number of indels implied before the i_m nucleotide, between i_m and j_m, and after j_m to the end. Because a centroid stem can be composed of multiple helices, not all the interval probabilities will necessarily be favorable. Thus, we look probabilities for all three intervals, and eliminate any candidates which have poor scores of less than 50% for all of them. Any remaining candidate that is also a feature is then included in the cluster.

5.4.5 Step five: Cluster validation

Each cluster passes through a final validation step; clusters that have broad support across sequences are validated, while clusters having support in only a few sequences are deleted. Support from each sequence is determined by the number of base pairs in the MFE structure of the search window. This represents the best, most energetically favorable scenario. If the number of MFE base pairs in a sequence's window is less than the centroid, support from the sequence is considered weak.

A cluster is validated with the following steps:

1. Set the total score T to the number of sequences in the initial cluster before refinement
2. For each missing sequence from the initial cluster, run MFE folding on the window defined in Section 5.4.4.2
3. Count the number base pairs g in the MFE substructure
4. Score the sequence by comparing g with the number of base pairs in the cluster centroid

5. Add each sequence's score to T

6. If total score T is above zero, keep the cluster; else, delete the cluster

The base pairings in a window is assesed by running minimum free energy (MFE) calculations on the window of interest, with constraints that the 5' and 3' ends cannot base pair with themselves. The sequence is assigned a score of 1 if the resulting minimum free energy structure has at least the number of base pairs as the centroid; a score of 0 if it has less base pairs than the centroid but at least half; a score of -1 if it contains less than half the number of base pairs as the centroid; and a score of -2 if the MFE structure has zero energetically favorable base pairs. Sequence scores are summed to assess overall support for the cluster, with any cluster with a zero or negative score eliminated as unfavorable.

5.4.6 Resampling

To further locate missing FN features, the sequences are resampled with constraints based on the clusters found thus far. Clusters with broad sequence support define an initial set of features from every sequence that can tentatively be considered TP. The features from each sequence that are not part of a cluster, then, can be considered FP: predicted by Boltzmann sampling, but not backed up by enough structural support across all sequences to be significant. Since these FP features preclude other, potentially native helices from forming, the sequence is resampled again while prohibiting the FP features. The basic steps in resampling are:

1. Form constraints file forbidding all features not contained in cluster by end of cluster validation

2. If the first iteration, resample each sequence using the constraints file

3. If not the first iteration, check whether any clusters are not present in previous iteration; if so, resample with constraints file

4. If no new clusters are formed between the present iteration and the last, terminate the program and output the final list of clusters

However, prohibiting the FP features could result in a sample significantly less energetically favorable than the original, to the point of being implausible. Hence, if the new sample's MFE structure is more than one standard deviation below from the median free energy of the old, then the new sample is considered implausible and not used, with the original sample being employed instead.

The entire algorithm is run again after Boltzmann sampling with constraints: profiling, clustering, refinement, and validation. If, after the second iteration, any new clusters are found, then the resampling occurs again, incorporating the data from the new clusters. If no new clusters are found, then the process terminates, with the last set of clusters presented as the final output. Termination is guaranteed, because all previously found TP features are retained; either the shrinking set of 'new' features or the increasing free energy suboptimality will limit the number of new clusters to be discovered.

At termination, a list of clusters in outputted with its set of sequence/feature pairs, and its representative centroid stem (i_c, j_c, k_c, l_c). In the tRNA example, `ConsensusStems` terminates with four clusters having 30, 27, 29, and 28 sequence/feature pairs, with respective centroid stems: $(1, 71.5, 7, 7)$, $(10, 25, 4, 4)$, $(22, 48, 9, 10)$, and $(47.5, 65, 6, 6.5)$. These centroids are very close to the Rfam consensus stems of $(1, 70, 7, 7)$, $(10, 24, 4, 4)$, $(26, 42, 5, 5)$, and $(48, 64, 5, 5)$. Indeed, we shall see that the high accuracy of `ConsensusStems` in predicting the native stems is reflected across all tested Rfam families.

5.5 Results

Given a set of homologous RNA sequences as input, `ConsensusStems` outputs a list of clusters denoted by their centroids. These $[i_c, j_c, k_c, l_c]$ quadruples are the consensus stem predictions. The sequence/feature pairs associated with each cluster indicate the support for this prediction across the family. Hence, we evaluate consensus prediction accuracy at three levels of structural granularity: base pair, stem, and centroid. Additionally, the dependence on number of sequences in the test family is analyzed.

Accuracy is measured by precision, recall, and Mathews Correlation Coefficient (MCC) [5]. These depend on the number of TP, FP, and FN structural elements, denoted tp, fp, and

fn. Precision is calculated as P = $\frac{tp}{tp+fp}$ and recall as R = $\frac{tp}{tp+fn}$ while MCC is approximated [46, 53] as their geometric mean: MCC $\approx \sqrt{PR}$. Hence, to evaulated accuracy, we must define, and then count, TP, FP, and FN at each level of granularity. To illustrate, we return to our initial tRNA *T.brucei* example.

5.5.1 At the base pair level

Although `ConsensusStems` harnesses the power of structural abstraction [132], base pair accuracies are measured for two reasons. First, this is the usual standard [35], so confirms that this new approach does no worse than other methods. Second, and more importantly, the contrast in accuraccies validates the choice of structural granularity. Base pairing interactions which are perceived as contradictory become a unified pattern when consolidated into a single stem, i.e. extended helix. Thus, the *exact same data* at higher abstraction/lower granularity is a much stronger and more accurate structural signal.

A predicted pairing is a TP if it appears in the Rfam alignment for that sequence. However [46], it is only classified as a FP if it actively contradicts the Rfam structure (although still counted in the prediction size). A canonical, noncrossing native base pair is a FP if it is not predicted; Boltzmann sampling does not predict noncanonical and/or pseudoknotted pairings so these are not counted.

To generate predicted base pairs for a given sequence, each feature (i, j, k) associated with a cluster is decomposed as $(i, j), \ldots, (i + k - 1, j - k + 1)$. *T.brucei* has two features, $(46, 66, 7)$ and $(50, 60, 4)$, associated with the same cluster output by `ConsensusStems`. The first contains the native helix $(48, 64, 5)$ from the Rfam alignment for *T.brucei*. Hence, for this cluster and this sequence, $tp = 5$, $fp = 2 + 4$, and $fn = 0$.

The base pair accuracy for the 11 Rfam test families is listed in Table 9. It was calculated by summing over each cluster and each sequence. Hence, the precision denominator is the total number of predicted base pairs, with multiplicity, across the entire family. These accuracies are comparable to other state-of-the-art consensus prediction methods [173, 4, 101].

Family	Precision	Recall	MCC
tRNA+	0.52	0.76	0.63
THF	0.54	0.91	0.70
TPP*	0.44	0.74	0.57
5S*+	0.50	0.76	0.62
FMN	0.28	0.78	0.47
U1	0.31	0.58	0.42
ykoK+	0.52	0.80	0.65
glmS+	0.41	0.84	0.58
IRES cripavirus	0.35	0.65	0.48
IRES HCV*	0.25	0.46	0.34
metazoa SRP	0.58	0.73	0.65
Avg	0.43	0.73	0.56
Stdev	0.11	0.13	0.11

Table 9: Base pair accuracy as described in Section 5.5.1. Values are comparable to other consensus methods. Note the low average precision relative to recall.

Family	Precision	Recall	MCC
tRNA+	0.97	0.91	0.94
THF	0.96	0.96	0.96
TPP*	0.98	0.84	0.91
5S*+	0.99	0.98	0.98
FMN	0.98	0.88	0.93
U1	0.95	0.90	0.93
ykoK+	1.00	0.97	0.98
glmS+	0.80	0.85	0.82
IRES cripavirus	0.83	0.85	0.84
IRES HCV*	0.99	0.81	0.90
metazoa SRP	0.91	0.71	0.80
Avg	0.94	0.88	0.91
Stdev	0.07	0.08	0.06

Table 10: Stem accuracy according to Section 5.5.2. Predictions, especially precision, have improved measurably with the reduction in granularity.

5.5.2 At the stem level

The *T.brucei* helices $(46, 66, 7)$ and $(50, 60, 4)$ are in conflict at the base pair level, since their coordinates overlap. However, they reinforce a clear, common structural signal at a higher level of abstraction — that the 5' and 3' regions of the stem $(46, 66, 8, 10)$ interact.

To consolidate this information, all features for a given sequence in a particular cluster are grouped under a single stem with coordinates (i, j, k, l). This indicates that all cluster pairings (i', j') for this sequence have endpoints with $i \leq i' \leq i+k-1$ and $j-l+1 \leq j' \leq j$ for the shortest possible segments. Hence, the stem $(46, 66, 8, 10)$ communicates that regions $[46, \ldots, 53]$ and $[57, \ldots, 66]$ of the *T.brucei* sequence interact, although the specific base pairings may belong to either feature $(46, 66, 7)$ or $(50, 60, 4)$, or even some other helix.

This abstraction also addresses the situation when one helix extends another in close enough succession to be clustered together.

A predicted stem is a TP if it intersects at least 50% of both 5' and 3' regions of a native stem [76, 4]. Otherwise, it is a FP. As with base pairs, predicted stems that are not in the native but do not contradict it are excluded from being counted as a FP; at least 50% of the stem must not be contradictory for this rule to apply. A native stem which is not so intersected by a predicted one is a FN. According to the Rfam alignment, native stems for *T.brucei* include $(48, 64, 5, 5)$. Since this is a subset of $(46, 66, 8, 10)$, the prediction is a TP.

The stem accuracy of each family, listed in Table 10, was calculated like base pairs, by summing over each cluster and each sequence. At this level of granularity, the wisdom of not trying to resolve competing base pairings is clear. Instead, those signals have been consolidated into a single, coherent regional interaction, and the resulting increase in accuracy, especially in precision as illustrated in Figure 31, is substantial.

5.5.3 At the centroid level

The previous section evaluated prediction accuracy for a family according to the pooled sequences' stem accuracies. We now consider the consensus stem predictions; recall that the cluster centroid (i_c, j_c, k_c, l_c) is the median of its constituent sequences' (i, j, k, l) stem coordinates.

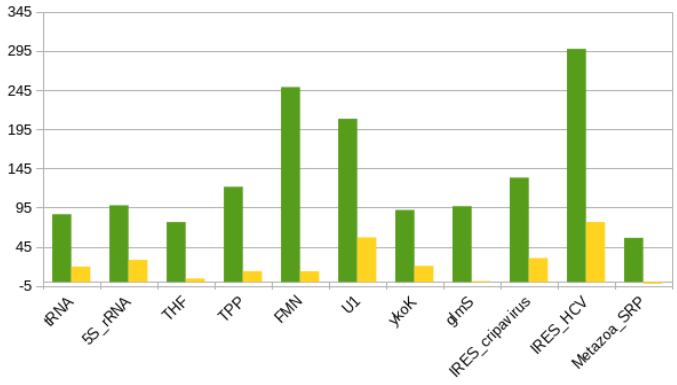


Figure 31: Percentage accuracy increases with increasing abstraction from base pairs to stems for precision (green) and recall (yellow). While detailed prediction remains difficult, a clear structural signal emerges at the higher level of structural abstraction.

The native consensus stems were determined from the Rfam consensus structure for each family. A quadruple (i, j, k, l) is the maximal set such that all base pairs (i', j') are located within the stem ($i \leq i' \leq i + k - 1$ and $j - l + 1 \leq j' \leq j$), are nested within each other (if $i_A < i_B$, then $j_B < j_A$), and with no non-nested base pairs (i_n, j_n) occurring such that either i_n or j_n is located in the regions $[i, i + k - 1]$ or $[j - l + 1, j]$. Visually, this is represented in the nested parentheses found in final line of the Rfam consensus alignment as an (i, j, k, l) region containing base pairs of the same symbol: '{}', '[]', '¡¿', or '()'.

Rfam's tRNA consensus structure includes the stem $(47, 63, 5, 5)$ while `ConsensusStems` outputs the centroid $(47.5, 65, 6, 6.5)$ for the cluster which contains the *T.brucei* stem $(46, 66, 8, 10)$. The cluster centroid is a TP since it overlaps with more than 50% of a native consensus stem. If no centroid so overlapped, the native would be a FN. If the centroid did not overlap with any native (or overlapped a native by less than 50%), it would be a FP.

The results in Table 11 show that 7 of 11 families tested have perfect consensus stem prediction accuracy. Average recall increased slightly over the summed sequence values in Table 10, while average precision remained essentially constant. The increase in recall is consistent with the value of consensus structure prediction; homologous sequences compensate for FN predictions.

The fact that accuracy values can drop as well as increase reflects the very different

Family	Precision	Recall	MCC
tRNA+	1.00	1.00	1.00
THF	1.00	1.00	1.00
TPP*	1.00	1.00	1.00
5S*+	1.00	1.00	1.00
FMN	1.00	0.80	0.89
U1	1.00	1.00	1.00
ykoK+	1.00	1.00	1.00
glmS+	0.75	0.75	0.75
IRES cripavirus	0.80	1.00	0.89
IRES HCV*	1.00	1.00	1.00
metazoa SRP	0.86	0.86	0.86
Avg	0.95	0.95	0.94
Stdev	0.08	0.10	0.08

Table 11: Cluster centroid accuracy as in Section 5.5.3. At this scale, the method has perfect precision and recall for 64% of the test families.

numbers of predictions when moving from the sequence/stem pairs to cluster centroids. For example, the glmS family has 77 TP, 15 FP, and 11 FN stem predictions but only 3 TP, 1 FP, and 1 FN centroids. Hence, the fraction of errors is higher for centroids since the set sizes are an order of magnitude smaller.

5.5.4 Tests with reduced sequence sets

The accuracy of `ConsensusStems` does depend on a large enough pool of Boltzman samples. This effect was assessed by running six additional trials per family: three sets of size 15 and then three of size 8 randomly chosen sequences from the original pool (summarized in Table 8).

Dropping the set size to 15 decreases the average centroid precision, recall and MCC by -2.6%, 10.8%, and 5.1%. Reducing the number further to 8 results in average decreases of 0.9%, 21.3%, and 13.8% respectively.

The corresponding changes at the base pair level were 3.5%, 38.4%, and 23.1% and then 3.8%, 67.1%, and 44.0%. Unsurprisingly, recall is affected much more severely than precision by decreases in the initial amount of information. This is especially apparent at high granularity.

5.6 Discussion

Results demonstrate that Boltzmann sampling of homologous sequences filtered by RNA profiling and noise-sensitive clustering resolves the consensus structure problem at the stem level of granularity. Namely, `ConsensusStems` output clusters recover the native regions of interaction with a high degree of accuracy on average, and low standard deviations, for a comprehensive test set of 11 Rfam families.

The cluster centroids, which are the consensus stem predictions for the entire family, have an average (approximate) Mathews Correlation Coefficient (MCC) of 95%, with an 8% standard deviation. MCC is useful as a summary statistic since it incorporate both precision (how many predicted elements are native) and recall (how may native elements are predicted) into a single value. The associated individual sequence stems, which are the 'alignment' produced by this method, similarly have an average MCC of 91% with a 6% standard deviation. In comparison, the base pair level MCC average is just 56% with std = 11%.

Improvements in RNA prediction accuracy achieved by structural elements at lower granularity than base pairs is a known phenomenon [132]. Here, profiling's methodology has been extended from maximal helices to stems, a higher level of structural abstraction. Stems consolidate competing helices often predicted by thermodynamic optimization [35] into a single coherent structural signal — two regions interacting with high probability. By turning 'competitors' into 'allies', `ConsensusStems` achieves high accuracy, correctly predicting the forest by not trying to resolve each tree.

The power of abstraction is such that resonable consensus predictions can be achieved on relatively small sets of sequences. That recall suffers disproportionately is not surprising; the method extracts the signal present with high precision, but cannot consolidate what is not there. Hence, sufficiently many homologous sequences (26 on average for the 11 Rfam families tested) are needed for full recall.

In comparison to other stem-based consensus prediction approaches, `ConsensusStems` has been more comprehesively tested [76] and more rigorously evaluated [4]. The `comRNA` program [76] identifies conserved helices in unaligned sequences using a graph-theoretic

approach. Three familes were tested, with average precision of 86.7% but no reported recall. The longest sequences were ∼200 nucleotides (nt), and most number of native stems were 5. Here, 11 families were tested with sequence lengths up to 300 nt, and up to 10 native consensus stems. The `RNAscf` [4] program achieved average precision and recall of 88.4% and 92.6% respectively over 12 test families with sequence lengths up to 200 nt and up to 5 stems. However, these values counted any overlap between predicted and native stems as a true positive, not the 50% minimum required here.

Moreover, while others [76, 4] perform exhaustive searches to find all possible helices, `ConsensusStems` harnesses the power of Boltzmann sampling to generate only the most probable ones. Their analysis is then very fast; RNA profiling extracts features from a set of structures in time linear in their size [134], and the noise-sensitive clustering algorithm DBSCAN has an average complexity of $O(n \log n)$ [43]. While the cubic runtime of Boltzmann sampling [32] is the bottleneck here, it is orders of magnitude lower than many consensus prediction methods [92, 102]. Additionally, if speed-ups are desired, it would be straight-forward to parallelize the sampling component of this approach since the individual input sequences have no data dependencies. Thus, `ConsensusStems` embodies the best of both worlds, achieving high accuracy in an efficient manner.

5.7 Conclusion

Predicting a common structure for a family of RNA sequences is is an old and important open problem in computational biology. It is challenging in no small part because RNA structures are much more strongly conserved than sequence identity. Our new `ConsensusStems` method addresses this challenge by (1) finding the right balance between precision and recall, and (2) selecting on an appropriate level of structural granularity.

First, Boltzmann sampling of sufficiently many homologous sequences eliminates nearly all false negative predictions, while noise-sensitive clustering of RNA profiling features filters almost all false positive ones. Second, focusing on the 'forest' of interacting sequence segments, rather than the 'trees' of specific base pairs, yields clear and accurate predictions, both for the cluster centroids as consensus stems as well as the supporting sequence/feature

pairs across the family.

Thus, our method succeeds in predicting the consensus stems with a high, often perfect degree of accuracy, as tested on a diverse group of families whose lengths span the range where the thermodynamic model is the most accurate. Even if a finer grained consensus prediction is desired, `ConsensusStems` should be used to make the initial lower granularity prediction. The predicted consensus stems can then be fine tuned, either by applying sequence alignment tools [42] to the features of the cluster, or by using the predicted stems as the known structural input to an alignment method [99].

Hence, both on its own and as a initial step toward accurate base pair level prediction, `ConsensusStems` represents a significant advance to the state-of-the-art of consensus structure prediction.

5.8 Funding

Funding for this paper was provided in part by the Burroughs Wellcome Fund [CASI #1005094 to C.E.H.], and by the NIH [R01 6M126554].

CHAPTER VI

DNA MIXTURE STUDY: QUANTIFYING THE INTRA- AND INTER-LABORATORY VARIABILITY IN FORENSICE DNA MIXTURE INTERPRETATION

6.1 Abstract

Despite the prevalence and weight of forensic DNA evidence in the criminal justice system, little is known concerning the variability in forensic DNA interpretation quality. Variability in interpretation is especially likely when the DNA sample has complicating factors. One major such factor is when more than one source of DNA is present in the sample, resulting in a DNA mixture. We present the first ever wide scale quantitative assessment of interpretation variability in forensic DNA mixture interpretation. We introduce novel metrics to measure the accuracy and precision of interpretation. Results of applying these metrics to the interpretations demonstrate: 1) a significant amount of variability exist both within and between laboratories; 2) that high quality interpretations are possible, with accuracy and precision being highly correlated. These point to the ongoing need for training and benchmarking within laboratories, and the need for dissemination of best practices between laboratories.

6.2 Introduction

Considered the reliable standard in forensics, DNA analysis carries the connotation of hard science and hence infallibility. DNA evidence often plays a significant role in either convicting or exonerating persons of interest. Thus, the accuracy and precision of DNA forensic analysis is essential.

Although the science behind DNA profile generation is reliable and repeatable, the interpretation of this data is not completely free of subjectivity. Previous DNA mixture studies by the National Institute of Standards and Technology (NIST) have shown variability when the same DNA mixtures were submitted to multiple laboratories [40] (NIST

Study 2005 and 2013; https://strbase.nist.gov/). This variability may be compounded as the complexity of a DNA sample increases, but the degree of variability present in DNA mixture interpretation by the forensic community is currently unknown. Thus, the size and the acceptable limits of variability by the forensic DNA community is also unknown. It is important to note that variability does not necessarily imply that an incorrect genotype was generated, but that the reported genotypes may include extraneous genotypes that differ among examiners interpreting the same data.

The purpose of this study was to assess the current state of DNA mixture interpretation in the forensic DNA community. Specifically, this study investigates the variability in the precision and accuracy of DNA examiners' DNA mixture interpretations given the same DNA *.fsa* files. While other DNA mixture studies have been conducted, results have been reported on a broad, mainly qualitative level. Hence, the results of the study are presented as follows: 1) we developed novel metrics to quantify a DNA exmainer's accuracy and precision in interpreting a varitety of DNA mixtures and 2) we use these novel metrics to determine the current variability range within the forensic DNA community with 2- and 3-person DNA mixtures.

The amount of variation that exists, and whether that variation is consistent within and between laboratories, is of interest to the forensic DNA community. Because DNA training and interpretation protocols are determined by each individual laboratory, we investigate whether intra-laboratory variability, where examiners utilize identical protocols and training, will be significantly different than inter-laboratory variability, where protocol and training differences are expected. The metrics developed by the study will also provide insight into strengthing and improving the current state of forensic DNA training and quality control. The quantitative data and novel metrics can be used to benchmark an examiners interpretation performance, determine mixture interpretation limitations within a laboratory, and infer whether a new method implemented in a laboratory yields improved precision and accuracy over previous methodologies.

6.3 Background/Related works

Current forensic analysis of DNA relies on sections of noncoding DNA, composed of small repeating fragments, usually 100-450 nucleotides in length. Known as short tandem repeats (STRs), these repeats occur at multiple locations in the genome and forensics utilizes a select few to generate a genetic profile. The exact number of STRs varies widely enough between individuals and, when multiple locations (loci) of repeats are considered in combination, they can be used to discriminate one person from another. During analysis of a DNA sample contaning a single contributor, the genetic profile can be determined relatively easily. When additional contributors are added to a sample, the complexity of the sample is increased and it may be difficult to distinguish which repeat belongs to each particular contributor.

Extensive research and evaluation has gone into refining the data interpreation of the generated STR data. With the popularity of DNA in pop culture (television forensics and courtroom dramas), the presence and absence of DNA forensic analysis can play a leading factor in determining a verdict [77]. Despite the apparent objective nature of DNA evidence, results are influenced by the ability of its practitioners to accurately interpret the results and in a manner that can duplicated by another DNA examiner.

Laboratory accreditation is intended to address some of these issues by implementing quality controls and establishing quality assurance systems to theoretically minimize error and improve consistency. The FBI has generated DNA guidelines with widespread adoption (SWGDAM 2010 guidelines), but the minute interpretation guidelines and limits are largely set by each laboratory. The quality of interpretation and execution is also influenced by examiner skillset and the quality of the DNA data generated from a given sample. Since individual laboratories determine their DNA protocols and interpretaion guidelines (reagents used, DNA amplification kits chosen, analytical and stochastich thresholds determined for the data [18], the role of known contributors in analysis [18]), [67] inter-laboratory interpretation results are likely to vary between laboratories. Likewise, variability between examiners within a laboratory may exist, due to human interpretation bias [18, 17, 38, 37, 81, 88].

Sample quality presents its own challenges to interpretation citepaoletti-etal-05. Factors such as low levels of DNA template [6, 52, 83, 9], can negatively impact interpretation

complexity and may contribute to the interpretation variation. In addition, determining the exact number of contributors (NOC) in a mixture is a non-trivial problem [91, 120, 16]; ambiguity in determining the NOC increases as the NOC increases in a sample and it is challenging enough that computational programs have been developed to address this problem [123]. Stochastic effects can also complicate interpretation of DNA data [122, 91], and the increase in allelic dropout [128, 126] or stutter [97, 12, 13, 14]), can also increase the complexity of the data.

Overlapping genotypes between sources can also complicate analysis [155, 23]; this problem is increased if the contributors are genetically related, making separation of the data more difficult. Ease of analysis is also related to the ratio of DNA sources in a mixture; evidence suggests that imbalanced ratios increase the chance for allelic dropout in the minor profile [51], while balanced contributor ratios present their own challenges to separate genotypes for each contriburtor with ambiguous data (RFU peak heights that are relatively equivalent).

The difficulty of mixture interpretation is compounded by the lack of consensus regarding standard methods and protocols for analysis and interpretation of DNA mixtures [18, 124, 125], the use of quantitative vs qualitative methods [124], the type of statistic applied, and the role of software and computational programs. Thus, the state of DNA mixture interpretation with respect to its accuracy and precision remains an important open question. Past inter-laboratory studies include those conducted by the National Institute of Standards and Technology (NIST) [86, 39, 87]. Similar collaborative studies have also been carried out by the European DNA Profiling Group (EDNAP), with qualitative differences being reported. Blind trials testing multiple laboratories have also been performed in other countries, notably by the German DNA Profiling Group (GEDNAP [130, 129]). Results from these previous studies, however, have focused on general trends and qualitative assessments, with reports of "results obtained by the vast majority of participating laboratories who consistently and reproducibly produce correct results" [129]. (GEDNAP studies have also identified the main source of errors as human carelessness [129], manifested in transcriptional and transpositional mistakes that this study has eliminated in order to uncover

deeper systematical errors.

Thus, by employing novel metrics, this study attempts to quantitatively identify the current variation in DNA mixture interpretation and represents an important contribution to the DNA mixture interpretation practices and the DNA forensic community.

6.4 Materials

The study was conducted by preparing a set of six samples, with four being a mixture of two DNA sources, and the remaining two being a mixture of three DNA sources. Each sample was analyzed, and the electropherogram files obtained. These files, along with a questionnaire and standardized worksheets to record interpretations, were then sent to forensic laboratories primarily from the United States for voluntary participation in the study. Over fifty laboratories with 185 examiners returned completed questionnaires and worksheets. This paper concerns the analysis of those questionnaires and worksheets.

6.4.1 Preparation of samples

Samples were taken from buccal swabs of 14 individuals, incubated at 56°C for 24 hours, extracted, and purified with Qiagen BioRobot EZ1 Advanced/Advanced XL with Investigator Card. The estimated concentration of DNA present from each contributor sample was determined by quantifying with the Applied Biosystems Plexor®HY quantification kit and 7500 HID instrument. Target DNA quantities were based on the male : human DNA ratio, with a target of 1 ng/uL DNA. The sample DNA was then amplified using the Applied Biosystems Geneamp^{TM} PCR System 9700 and separated by capillary electrophoresis on the Applied Biosystems®3130xl Genetic Analyzer. Single source profiles were generated for each of the 14 contributors in order to serve as a key for the mixtures.

The 14 individual profiles were populated into NIST's Virtual Mixture Maker (https://strbase.nist.gov /software.htm) to develop hypothetical 2- and 3-person mixtures. The program performs a pairwise comparison of STR profiles in a datasest and calculates the number of loci possessing 1-6 alleles in all possible mixtures. (This program was also used in the NIST Interlaboratory Mixture Interpretation Study 2005 [MIX05]). Comparisons were made across all possible mixtures and the median allelic overlap from the 2- and 3-person mixtures were

	Qty (ng/uL)	Vol (uL)	TE (uL)
Source 18	6.0	8.3	491.7
Source 21	0.1	NA	NA
Source 24	2.7	18.5	481.5
Source 25	16.0	3.2	496.8
Source 31	6.3	7.9	492.1
Source 35	4.9	10.2	489.8
Source 39	9.5	5.3	494.7
Source 44	4.8	10.4	489.6
Source 53	4.1	12.2	487.8
Source 55	3.3	15.2	484.8
Source 57	22.0	2.3	497.7
Source 60	2.6	19.3	480.7
Source 62	0.8	13.3	86.7
Source 64	3.1	16.2	483.8

Table 12: Table R, with values for estimated quantitation, sample volume, and TE buffer volume. All single source sample, except Samples 21 and 62, were normalized to 0.100ng/uL in a final volume ot 500uL of TE buffer, by using the concentration values from the quantification data in the formula $(C_1)(V_1) = (C_2)(V_2)$. Samples 21 and 62 could not be normalized due to their low quantitation values; however, they were still used in creating the mixtures.

Sample	2-person			3-person	
	2:1	3:1	4:1	4:1:1	1:1:1
18 & 31	–	–	80+20	–	–
64 & 21	50+25	–	–	–	–
55 & 35	50+25	–	–	–	–
44 & 62	–	75+25	–	–	–
53, 60 & 25	–	–	–	40+10+10	–
57, 24 & 39	–	–	–	–	25+25+25

Table 13: Table S, detailing the mixture compositions, with their volumes in uL forming either a 2:1, 3:1, 4:1, 4:1:1 or 1:1:1 ratio. A total of seven mixtures were generated, four 2-person mixtures and two 3-person mixtures. All mixtures were quanted with Plexor®HY, amplified in triplicate, and analyzed on the 3130XL CE Genetic Analyzer.

selected for the study. The individual profiles in the median allelic overlap were then used in the mixture generation as follows.

Accurate assessment of the single source sample concentrations allowed for the appropriate selection of DNA target quantities to be used in order to generate mixtures at the desired ratios. All single source samples (except Samples 21 and 62) were normalized to 0.100ng/uL in a final volume of 500uL of Te buffer (see Table 12). This was done by using the concentration values from the quantification data and following the formula $(C_1)(V_1) = (C_2)(V_2)$, where C and V refer to the concentration and volumes respectively. Samples 21 and 62 could not be normalized due to their estimated quantitation values; however, they were still

used in creating the mixtures and adjusted via peak height ratios round after analysis in the Applied Biosystems®3130xl Genetic Analyzer.

Two and 3-person mixtures were generated using Identifiler®Plus and PowerPlex®16 HS amplicfication kits. Single source samples were amplified at a target of 0.5ng with the PowerPlex®16 kit while the mixture samples were amplified at a target of 0.7ng with the PowerPlex®16 kit and 0.4ng with the PowerPlex®16 HS kit. The 3130xL CE instrument was calibrated for each respective kit by running the matrix standards according to the kit manufacturer's instructions. The resulting .fsa files were analyzed utilizing the GeneMapper®IDX software (version 1.0.1/1.1) and eletropherograms for each sample were produced.

A total of six mixtures were generated, four of which were 2-person mixtures and two 3-person mixtures. All mixtures were then quantified with Plexor®HY, followed by amplification in triplicate and analysis on the 3130xL CE Genetic Analyzer. Mixture ratios are listed in Table 13. Peak height response (intensity signal) was used to determine the ratio of one mixture to the other. All peak heights for a given contributor to a mixture were summed and then compared with the sum of peak heights from the other contributors to get a ratio of contributors. If the ratio needed to be adjusted for later runs, the concentrations were adjusted based on the peak height response.

6.4.2 Examiner participation

Participants in the study were solicited via forensic conference presentations, the *Crime Lab Minute* newsletter from the American Society of Crime Laboratory Directors, and direct solicitation to DNA Technical Leaders across American forensic laboratories. Each laboratory was provided a questionnaire, DNA mixture data, and a response worksheet to record their analysis. Each laboratory was requested to have each individual examiner complete the interpretation and submit their own interpretation worksheet.

Over fifty laboratories with 185 individual examiners responded, sending back completed interpretation worksheets. Laboratories came mainly from the U.S., with a few international labs participating from Canada, the United Kingdom and New Zealand. The U.S. labs were

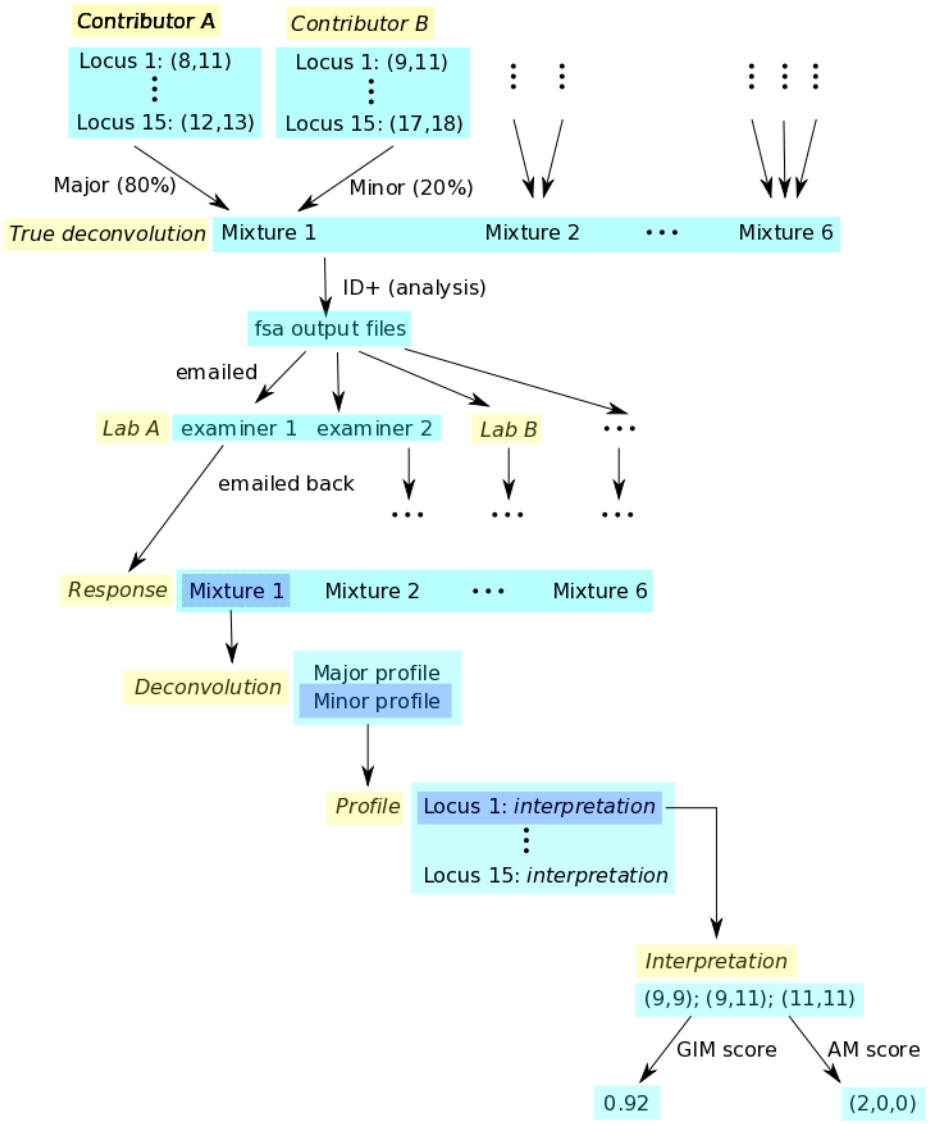


Figure 32: Schematic detailing a high level view of the process from making the sample through sending it to examiners to scoring the returned interpretation

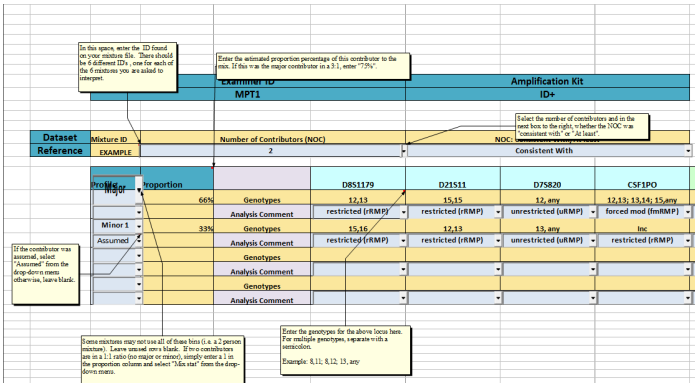


Figure 33: Screenshot of a sample spreadsheet with instructions given to examiners to fill with their interpretation.

categorized as either local (i.e. affiliated with a city), state or federal, with local labs being the most numerous.

Examiners were sent the .fsa files from either the ID+ or PowerPlex16 kit based on the type kit routinely used by the lab. In addition to the six .fsa files, every examiner was sent an excel spreadsheet with instructions in which to enter their interpretation and comments. Fields were present in the spreadsheet to enter data for each mixture on its number of contributors (NOC), mixture proportion, interpretation or genotype for each locus in a profile, comments on the type of analysis performed on the locus, interpretation model used, final statistic and any other comments.

Also submitted to each examiner for completion was a survey concerning the examiner himself. Questions included the examiner's education level, years of experience, most influential factor in interpretation, formal training, caseload, quantitative length of time to interpret each mixture and qualitative assessment of the difficulty of each mixture.

6.5 Methods

The goals of this study are twofold:

1. Uncover the general state of DNA mixture deconvolution overall, to discover the limits of interpretation and areas of variability

2. Uncover the relative state of DNA mixture deconvolution per lab, to help further training and quality assurance

These two goals will be addressed in this paper by developing three things: metrics, analysis and visualization. Specifically, we develop metrics to quantify an interpretation's accuracy and precision. Based on these metrics, we analyze the distribution and variation of scores by calculating median scores and interquartile ranges (IQR). Insight can be gleaned from these scores by visualizing median and interquartile ranges (plus outliers) with non-parametric boxplots, which serves to compare distributions of scores between labs. We also use scatterplots to visualize tradeoffs between accuracy and precision, and to explore scores that err with too much or too little caution.

More specifically, we explore the first goal by uncovering both high scores and variability. Top accuracy scores indicate the current limit of interpretability, one that others can potentially be brought up to. Areas of tight agreement, as revealed by the IQR score, indicate consensus among the community of its interpretability, whereas mixtures with a wide range of answers indicate gray areas that need clarification in training and protocol, perhaps by sharing of best practices and/or protocol standardization.

The second, related goal can be achieved by grouping results by lab. Because each lab is responsible for its own training and quality assurance, it is instructive to measure results, and recommend measures for improvement, on a lab-by-lab basis. Two scenarios that point to a need for improved quality assurance can emerge when analyzing results by lab: 1) a lab could have a large spread of scores (as seen by a large IQR), indicating adherence to and/or clarification of protocols are needed to promote reproducibility within a lab; and 2) the lab could have a close agreement on an interpretation that, compared with other labs in the graphs, is not as good as it can be. In other words, while its IQR may be small, its overall median falls below others. This indicates that while adherence to the protocol is not an issue, the protocol itself may be too conservative as compared to the what is possible.

6.5.1 Metrics

Thus, in order to uncover high and low scoring examiners individually (and labs collectively), there must be a way to quantify and compare the effectiveness of an interpretation. Because interpretations are often complex and include a range of possibilities, the ideal high quality

mixture interpretation has at least two characteristics: it identifies the correct genotype, and excludes as many incorrect ones as possible. These two characteristics can loosely be called "accuracy" and "precision", two qualities that have previously been identified as crucial to a high quality interpretation [129] . (Although each allele pair in an interpretation is not strictly speaking an independent trial, we nevertheless use the terms "accuracy" and "precision" in a general, not scientific, sense to describe our metrics.) In other words, an ideal interpretation is both accurate in giving the correct answer, and precise in only giving the correct answer. In practice, interpretations can vary widely in both characteristics.

A two-pronged metric called the DNA Examiner Assessment Tool (DEAT) was developed to quantify both characteristics. The Allelic Match (AM) score is solely concerned with accuracy, and the Genotype Interpretation Metric (GIM) score is solely concerned with precision. Taken together, these two complementary scores reveal inter- and intra- laboratory variation on the quality of DNA mixture interpretation, and provide a way to zero in on unusually low scoring results. They are also a more detailed way to provide labs and individual examiners specific feedback in potential training and benchmarking scenarios.

6.5.1.1 The Allelic Match score

The AM score measures an interpretations accuracy, and is broken down into three subscores: the Allelic Truth, Allelic False, and Inc scores. They respectively score the number alleles correctly and incorrectly interpreted, as well as the number labeled inconclusive.

In order to precisely define both the AM and the GIM score, we use the following definitions *Just doing ID+ for now.* In DNA forensic analysis, fifteen genomic loci are typed as comparison points and are denoted $L = \{D8S1179, D21S11, D7S820, CSF1PO, D3S1358, THO1, D13S317, D16S539, D2S1338, D19S433, vWA, TPOX, D18S51, D5S818, FGA\}$. We denote the set of possible alleles (covering at least 95% of the population) for a locus $l \in L$ as A_l. For example, $A_{TPOX} = \{i \mid 6 \leq i \leq 13\}$. We also define an augmented set of alleles as $A'_l = A_l \cup \text{`}any\text{'}$.

A *combination* c_l for a locus l is denoted as a tuple (a, b) where $a \in A_l$, $b \in A'_l$, e.g. $c_{TPOX} = (8, 11)$ or $(8, any)$. If $a, b \in A_l$, then a, b are ordered such that $a \leq b$. An

interpretation for a locus is a set of combinations given for a locus l $i_l \in A_l*$, where A_l* is the Kleene star on A_l. If $i_l = \epsilon$, it is interpreted as "inconclusive". An example for $l = TPOX$ is $i_{TPOX} = \{(8,8),(8,11),(11,11)\}$.

A *profile* P is a set of interpretations for all fifteen loci, i.e. $P = \{i_l \ \forall l \in L\}$, that describe a single individual's DNA. By definition, a DNA *mixture* M has more than one individuals DNA, and hence its *deconvolution* D_M is denoted as a set of profiles, with the number of profiles or contributors (NOC) determined by the examiner, i.e. $D_M = \{P_j \mid 1 <= j <= NOC\}$. Since every examiner E was given six mixtures to interpret, a full *response* $R_E = \{D_k \mid 1 <= k <= 6\}$ is a set of six mixture deconvolutions.

A mixture may have multiple deconvolutions, but only one is the true or correct answer that accurately reflects the DNA of the contributing individuals. We represent each *contributor* C_k as a set of alleles $C_k = \{c_l* \mid c_l * \text{is the correct combination } c_l, \forall l \in L\}$. Hence, a mixture M with N contributors has a true deconvolution $T_M = \{C_k \mid 1 \leq k \leq N\}$. An examiners NOC is correct if NOC $= |T_M|$. Note that instead of having multiple combinations or INC possible as with other profiles, a contributor has only one combination per locus.

We now can define the allelic match score AM for each interpretation i of a locus as a tuple of three subscores: Allelic True (AT), Allelic False (AF), and Inconclusive (INC), i.e. $AM_i = (AT_i, AF_i, INC_i)$.

Given a mixture and a contributor, the true combination of a locus l is denoted $c*_l =$

$(x*, y*)$, where $x*, y* \in A_l$. We score a single combination $c_l = (x, y)$ as

$$AM(c_l) = (AT(c_l), AF(c_l), INC(c_l))$$

$$= \begin{cases} (2,0,0), & \text{if } (x = x*, y = y*) \\ (1,1,0), & \text{if } \begin{array}{l}(x* = x \text{ and } y! = y*) \text{ or } (x = y* \text{ and } y! = x*) \\ \quad \text{or } (y = x* \text{ and } x! = y*) \text{ or } (y = y* \text{ and } x! = x*)\end{array} \\ (1,0,0), & \text{if } (x = x* \text{ or } x = y*, y = \text{`}any') \\ (0,1,0), & \text{if } (x! = x* \text{ and } x! = y*, y = \text{`}any') \\ (0,2,0), & \text{if } (x! = x* \text{ and } x! = y* \text{ and } y! = x* \text{ and } y! = y*) \\ (0,0,2), & \text{if } c_l = \text{`}inc' \end{cases} \quad (2)$$

We define a total ordering on the tuple $AM = (AT, AF, INC)$ such that for two AM scores $AM_1 = (AT_1, AF_1, INC_1)$ and $AM_2 = (AT_2, AF_2, INC_2)$, $AM_1 < AM_2$ if $(AT_1 < AT_2)$ or if $(AT_1 = AT_2$ and $AF_1 > AF_2)$. In other words, an AM score is larger than another if its AT score is higher; if the AT scores are equal, then the higher AM score is the one with the smaller AF score. For an interpretation i_l, its Allelic Match score $AM(i_l)$ is calculated as the maximal score of all its combination, i.e. $AM(i_l) = AM(c'_l)$ such that $c'_l \in i_l$ and $\forall c_l \in i_l$, $AM(c_l) <= AM(c'_l)$.

Hence, we can assign an aggregate AM score for every profile, deconvolution and response by summing individual AM scores, applying vector addition to the AM tuple, i.e. for $AM_1 = (AT_1, AF_1, INC_1)$, $AM_2 = (AT_2, AF_2, INC_2)$, $AM_1 + AM_2 = (AT_1 + AT_2, AF_1 + AF_2, INC_1 + INC_2)$. For a profile P, its score is calculated $AM(P) = (AT(P), AF(P), INC(P)) = \sum_{l \in L} AM(i_l)$.The highest possible AT, AF or INC score for a profile of fifteen loci is hence thirty, since each locus is scored for its two alleles.

Similarly, for a deconvolution of a mixture D_M, its score is the sum of its component profiles' scores $AM(D_M) = (AT(D_M), AF(D_M), INC(D_M)) = \sum_{P_j \in D_M} AM(P_j)$, with the highest possible AT, AF or INC score being sixty for a two-person mixture, and ninety for a three-person mixture. For an entire six mixture response R, the total score is calculated as the sum of its component mixtures' scores $AM(R) = (AT(R), AF(R), INC(R)) = \sum D_M \text{ in } RAM(D_M)$.

6.5.1.2 The Genotype Interpretation Metric

While the AM score does not penalize the number of combinations included in a locus interpretation, clearly having less combinations is preferable to more. Hence, in addition to measuring accuracy with the AM score, we also measure precision with the Genotype Interpretation Metric (GIM) score.

The most precise interpretation is a single two-allele combination, which receives a perfect GIM score of 1. The least precise interpretation is the Inconclusive ("INC") label, which receives a GIM score of 0. For all other interpretations, the GIM score compares the number of combinations in the interpretation against the total number of combinations C_{str}, calculated from published allele frequency sets to cover 99.5%. Because a combination with an any' is much less precise than a two-allele combination, the GIM score penalizes the former more than the latter. Hence, given a non-INC locus interpretation $i_l = c_k$, we partition the set of k combinations into those containing an any' and those without:

$i_l = C_a \cup C_{wa}$, where $C_a = \{(a,b)|a \in A_l \text{ and } b = `any'\}$ and $C_{wa} = \{(a,b)|a,b \in A_l\}$. We measure its GIM score as

$$GIM(i_l) = \begin{cases} 1, & \text{if } C_a = \emptyset \text{ and } |C_{wa}| = 1 \\ 0, & \text{if } i_l = `INC' \\ \frac{1-\frac{|C_{wa}|}{C_{str}}}{2^{|C_a|}}, & \text{otherwise} \end{cases}$$

As with the AM score, we can assign an aggregate GIM score for every profile, deconvolution and response by summing individual GIM scores: the GIM score of a profile $GIM(P) = \sum_{l \in L} GIM(i_l)$ is the sum of its loci GIM scores, the score of a deconvolution $GIM(D_M) = \sum_{P_j \in D_M} GIM(P_j)$ is the sum of its profile scores, and the score of the entire response $GIM(R) = \sum_{D_M \in R} GIM(D_M)$ is the sum of all its mixture deconvolution scores.

6.5.2 Analytics

Having defined accuracy and precision metrics, we now can calculate statistics given a set of scores. These statistics should capture both the overall quality of a set of scores, and the variation or range contained therein. We measure the first by taking the median score

of the set, and the second by calculating the interquartile range of scores (the difference between the third and first quartile).

The median scores help achieve the first goal by uncovering those mixtures that are easily interpreted (high median) or clearly inconclusive (low median), as well as enabling us to track the effects of increasing mixture complexity. It also helps the second goal if a median score by a lab is significantly lower than other lab medians, or even the median overall score; improving a lab's median score is a clear and well-defined objective for improvement. The interquartile range, as a direct measure of variability, also highlights areas for improvement. Namely, it uncovers areas where protocols may either be ambiguous, or poorly enforced.

The set of scores on which to perform analytics can vary depending on the desired granularity of results. We group the data at various levels of granularity to expose any outliers or unusual degrees of variability: first by mixture, then by region, lab, profile and locus. Because this depth of analysis becomes cumbersome to repeat for all possible combinations, we pick representative ones to illustrate the degree of variation possible.

For a broader, more systematic study we focus on labs, as they are the context in which examiners are trained and tested, and are the most logical venue to implement any changes to protocol and quality control. Because we are also interested in intra-lab variability, only those labs with at least five examiners participating are analyzed.

6.5.3 Visualization

Having calculated median and IQR scores for different labs for accuracy and precision, it is instructive to visualize them with boxplots and scatterplots for the sake of comparison.

Boxplots directly visualize both the median score as the central red line, and the IQR as the top and bottom limits of the box. Boxplots allow easy visual inspection of both the overall quality and the variability for the labs' accuracy as well as precision. By plotting each lab's distribution, it also enables outliers to become readily apparent. Namely, those with a large box (IQR) or a low red line (median) are easy to spot and address.

While boxplots visualize median and IQR scores for either accuracy or precision among the labs, scatterplots can directly compare both. We plot either median or IQR scores of

accuracy against precision, with each lab represented as a dot whose size is proportional to lab size. This enables us to answer questions concerning another aspect of the state of mixture interpretation: is there necessarily a tradeoff between precision and accuracy? Do labs generally achieve one at the expense of the other? What is the ideal balance between the two?

6.6 Results

To satisfy the goals of 1) uncovering the limits of interpretation, and 2) identifying areas of improvement on a lab by lab basis, we examine not only all scores grouped together, but also scores grouped along different parameters.

As a preliminary exploration into the areas of interpretation variability, we analyze scores grouped by region (local, state, federal and international/other).While all mixtures show some variability, we chose to highlight Mixture 1, designed to be the easiest mixture with only two contributors, the largest targeted ratio of 3.5 to 1, and clear peak heights that average well over standard stochastic threshold. While the more difficult mixtures can be expected to have a spectrum of responses, it is instructive to delve into one of the easiest to receive a baseline view of variability.

We found that even with this baseline mixture, a noticeable amount of variability is found at all levels, both between and within groups. At the highest level, scores grouped by region show small differences in median score but larger ones in IQR (Figure 34 (a) and (b)). Federal labs have the best scores, with thehighest median score as well as the smallest IQR. Local labs show the most variation, with an IQR of slightly above 0.6 to slightly under 0.9 for both GIM and AT scores.

Delving further, we investigate whether the spread in scores is due to distinct differences between local labs, or is found within all labs. In order to investigate the variability within a lab, we examine the larger local labs using ID+ (the most popular amplification kit) with at least five examiners. Six such labs exist and are shown in Figure 34 (c) and (d). Both the GIM and AT scores indicate that the spread of scores is due to both variation between and within labs. In particular, the differences in median scores between labs is more striking

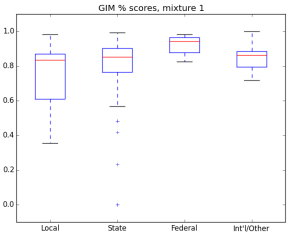
and pronounced than in regional scores, with median GIM and AT scores ranging from around 0.4 to well over 0.8. In terms of AT, this indicates that some labs (labs A, E and F) consistently achieve correct scores for over 80% of mixture 1 alleles, while another lab (lab D) consistently gets less than half correct. Additionally, while labs B and C have median scores between these two extremes, they also exhibit significant variation within their labs: their IQR's span from around 0.6 to almost 0.9.

Zeroing in on lab C, one of the labs with a larger range of scores within its eight examiners, we can investigate the nature of its range of scores by separating scores for the major versus minor contributor. Figure 34 (e) and (f) indicates that while there is a tight consensus in both GIM and AT scores for the major, the variability in overall mixture score comes primarily from the minor profile, whose IQR spans half (0.5) of the entire spectrum.

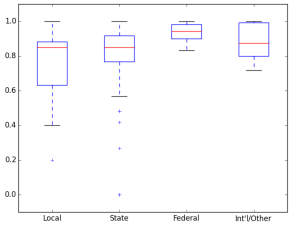
Finally, investigation of variability at the lowest level yields a loci-by-loci picture of the minor profile. Figure 34 (g) and (h) indicates that there is no consensus in either precision or accuracy on nearly all the loci. Since the GIM IQR includes a zero score on a majority of loci, this indicates more than one examiner marking the minor as 'inconclusive'. Yet since all but one loci has an AT IQR reaching 1, this also indicates more than one examiner correctly interpreting the minor profile.

Since examiner training and protocols are primarily the domain of each individual labs, variations in both median and IQR lab scores have actionable implications for the improvement of both. Namely, differences in median scores could be interpreted as differences in lab protocols. Similarly, differences in lab IQR scores are interpretable as differences in training or conformity to protocols. Figure 34 shows us that both exist, even at the simplest mixture level. Hence, this indicates that both in-house protocols and training need to be addressed.

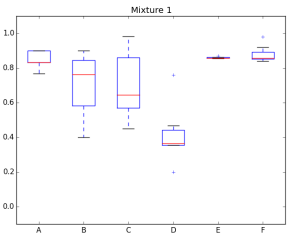
How widespread this variation is across mixtures is the next topic addressed. In order to assess the inter- and intra-laboratory variability of scores, we calculated the median and interquartile range (IQR) of the resultant AM and GIM scores by lab. To ensure appropriate sample size, we report only those labs with at least five examiners in Tables 1 and 2. For reference, we include statistics for all examiners in the thirteen large labs, in addition to all examiners from all labs.



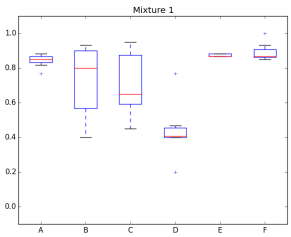
(a) Mixt. 1 by region, GIM



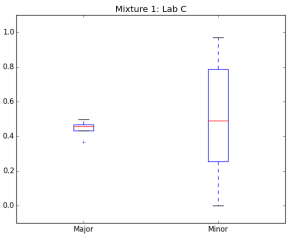
(b) Mixt. 1 by region, AT



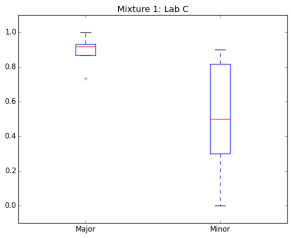
(c) Mixt. 1, large local labs, GIM



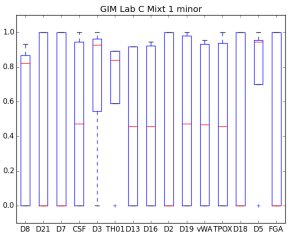
(d) Mixt. 1, large local labs, AT



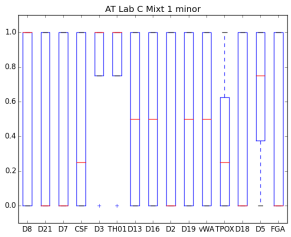
(e) Mixt. 1, lab C profiles, GIM



(f) Mixt. 1, lab C profiles, AT



(g) Mixt. 1, lab C, minor profile loci, GIM



(h) Mixt. 1, lab C, minor profile loci, AT

Figure 34: Preliminary data exploration of interpretation variability across regional, state, profile and loci levels. Each level, starting with regional at the top, occupies a row. The left column shows boxplots of GIM scores, while the right shows boxplots of AT scores. At every level, note the differences between boxes of both median scores (red line) and also interquartile range (box height), as well as the correlation between GIM and AT scores.

	Mixture 1		Mixture 2		Mixture 3		Mixture 4		Mixture 5		Mixture 6	
	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
Lab A	91.5	13.1	100.0	0.0	92.7	3.8	87.5	13.6	33.3	40.4	0.0	0.0
Lab B	83.3	6.7	100.0	6.7	30.4	11.5	76.7	7.6	59.5	35.1	29.7	34.2
Lab C	94.2	8.7	100.0	0.1	46.1	47.3	97.5	4.4	54.4	57.7	16.9	78.0
Lab D	90.0	9.7	86.7	3.3	0.0	0.0	78.3	11.7	33.3	0.0	0.0	0.0
Lab E	76.3	26.2	93.9	5.4	17.9	18.0	86.3	6.2	63.9	21.8	0.0	0.0
Lab F	64.5	29.2	91.9	5.1	18.0	1.4	69.9	2.1	20.0	11.1	0.0	0.0
Lab G	37.5	11.1	96.7	27.2	12.1	90.3	64.1	9.3	33.3	3.9	11.3	81.1
Lab H	82.8	11.7	100.0	8.0	0.0	29.8	71.7	34.1	33.3	4.6	0.0	0.0
Lab I	85.8	0.4	100.0	50.0	0.0	37.7	92.0	43.2	33.3	0.0	0.0	0.0
Lab J	70.0	7.4	93.3	13.3	82.8	0.0	52.7	35.8	21.7	12.0	17.4	83.3
Lab K	79.3	5.0	98.7	26.7	0.0	0.0	60.0	20.9	0.0	24.4	0.0	0.0
Lab L	85.8	4.1	100.0	0.0	7.6	33.5	92.3	19.8	47.9	7.2	0.0	0.0
Lab M	66.9	23.1	90.0	14.0	18.0	40.0	72.6	21.3	44.6	29.0	11.3	14.1
All big labs	84.5	16.5	98.3	11.7	18.0	59.8	78.3	22.0	33.3	24.8	0.0	11.3
All examiners	85.0	15.5	99.9	10.4	30.4	75.0	81.7	18.9	33.3	35.0	0.0	11.3

Table 14: Data for GIM scores in percentage of total possible for all labs with five or more examiners. Individual scores were computed per mixture for every examiner in the lab, and the overall median score and interquartile range (IQR) are reported. Median and IQR scores are reported for all combined examiners in a large lab, as well as all examiners in the study.

The range of complexity found in the six mixtures is reflected in the wide range of AM and GIM scores seen. To better understand the different scores given to examiners presented with the same *fsa* data file, we group the data per mixture. Because mixtures 5 and 6 are 3-person mixtures instead of 2-person (as is the case for mixtures 1-4), we normalize the $AT(D_M)$ score per mixture by converting it to a percentage of the total score possible. Namely, dividing $GIM(D_M)$ by 30 for a 2-person mixture and 45 for a 3-person mixture yields the percentage scores found in Table14. Normalized $AT(D_M)$ scores are similarly obtained by dividing by 60 and 90, respectively, as seen in Table 15

We compare the variability in precision i.e. GIM scores in the boxplots of Figures 35 and 36, and the variability in accuracy i.e. AT scores in the boxplots of Figures 35 and 36. Even for mixture 1, designed to be the easiest mixture with two contributors at 3.5:1 ratio, quite a spread exists both in the median lab score between labs, and in the range of scores found within a lab. This illustrates that the regional variations seen previously are not a fluke of grouping, but indicative of the general state of variability. A few labs have a median GIM score close to 100%, while others have a score below 70%, with one lab below 50%. Within labs, some labs have a tight range of scores between examiners, while others have an IQR of over 20%.

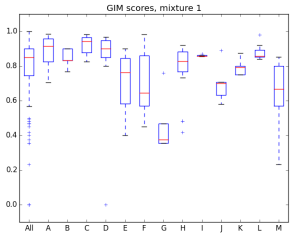
	Mixture 1		Mixture 2		Mixture 3		Mixture 4		Mixture 5		Mixture 6	
	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
Lab A	88.3	13.3	100.0	0.0	93.3	13.3	85.0	8.3	33.3	34.4	0.0	0.0
Lab B	85.0	3.3	100.0	6.7	30.0	6.7	75.0	5.0	60.0	34.4	28.9	28.9
Lab C	94.2	8.7	100.0	0.1	46.1	47.3	97.5	4.4	54.4	57.7	16.9	78.0
Lab D	90.0	9.7	86.7	3.3	0.0	0.0	78.3	11.7	33.3	0.0	0.0	0.0
Lab E	76.3	26.2	93.9	5.4	17.9	18.0	86.3	6.2	63.9	21.8	0.0	0.0
Lab F	64.5	29.2	91.9	5.1	18.0	1.4	69.9	2.1	20.0	11.1	0.0	0.0
Lab G	37.5	11.1	96.7	27.2	12.1	90.3	64.1	9.3	33.3	3.9	11.3	81.1
Lab H	82.8	11.7	100.0	8.0	0.0	29.8	71.7	34.1	33.3	4.6	0.0	0.0
Lab I	85.8	0.4	100.0	50.0	0.0	37.7	92.0	43.2	33.3	0.0	0.0	0.0
Lab J	70.0	7.4	93.3	13.3	82.8	0.0	52.7	35.8	21.7	12.0	17.4	83.3
Lab K	79.3	5.0	98.7	26.7	0.0	0.0	60.0	20.9	0.0	24.4	0.0	0.0
Lab L	85.8	4.1	100.0	0.0	7.6	33.5	92.3	19.8	47.9	7.2	0.0	0.0
Lab M	66.9	23.1	90.0	14.0	18.0	40.0	72.6	21.3	44.6	29.0	11.3	14.1
All big labs	84.5	16.5	98.3	11.7	18.0	59.8	78.3	22.0	33.3	24.8	0.0	11.3
All examiners	85.0	15.5	99.9	10.4	30.4	75.0	81.7	18.9	33.3	35.0	0.0	11.3

Table 15: Data for AT scores in percentage of total possible for all labs with five or more examiners. Individual scores were computed per mixture for every examiner in the lab, and the overall median score and interquartile range (IQR) are reported. Median and IQR scores are reported for all combined examiners in a large lab, as well as all examiners in the study.

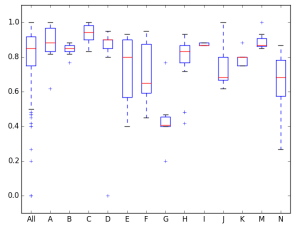
Interestingly, the general spread of AT scores resembles that of the GIM scores, with labs that scored high on precision also scoring well on accuracy, and similarly for low scoring labs. Additionally, the GIM IQR of each lab is also similar to its AT IQR, indicating that the amount of variability within a lab is consistent both with respect to accuracy as well as precision. Comparing GIM and AT median and IQR scores directly in Fig. 37 confirms this, with labs falling fairly close to the identity diagonal. More rigorous analysis using Spearman's coefficient shows a correlation factor of over 0.9 for almost all mixtures, for both median and IQR scores.

This patterns holds for mixture 2, a 2:1 mixture, with median scores considerably more uniform at or close to 100%. This is not unsurprising, given that a known profile was provided to examiners for this 2-person mixture. Even so, the range of scores found within labs still exhibits a significant spread, hitting 50% for one lab. As with mixture 1, labs having a large range of GIM scores also exhibit a large range of AT scores, indicating

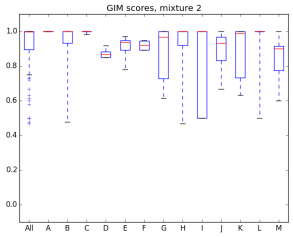
Although the correlation between GIM and AT scores persists in mixture 3, the scores for mixture 3 drop significantly in terms of both accuracy and precision. This indicates the general difficulty of making accurate and precise interpretations in the absence of a known profile. Nonetheless, some labs still exhibit a uniformly high level of accuracy and precision.



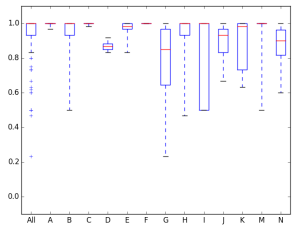
(a) Mixt. 1, GIM



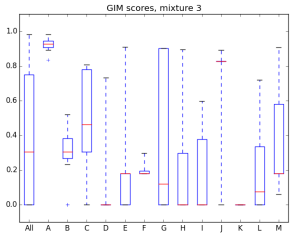
(b) Mixt. 1, AT



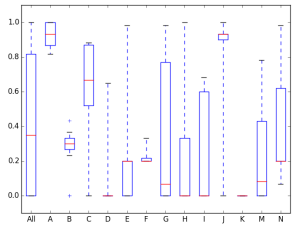
(c) Mixt. 2, GIM



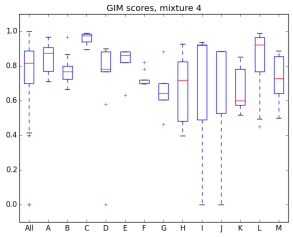
(d) Mixt. 2, AT



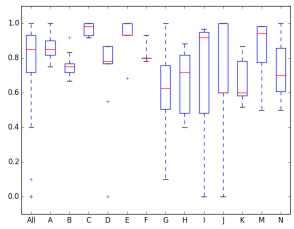
(e) Mixt. 3, GIM



(f) Mixt. 3, AT

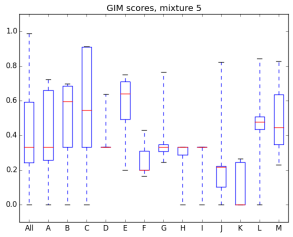


(g) Mixt. 4, GIM

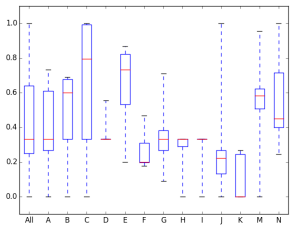


(h) Mixt. 4, AT

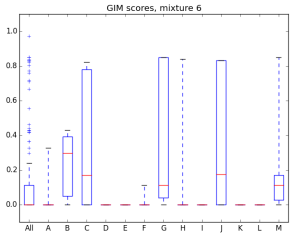
Figure 35: Boxplots for the 2-person mixtures $1-4$ of the thirteen labs of size five or greater, giving the distributions of each lab's respective examiners' scores. Red lines indicate median scores, boxes delimit the interquartile range, with outliers beyond it. The left column displays the GIM or precision scores from each lab, while the right column displays the AT or accuracy scores.



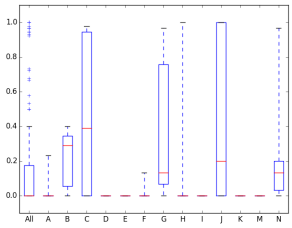
(a) Mixt. 5, GIM



(b) Mixt. 5, AT

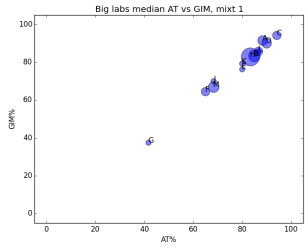


(c) Mixt. 6, GIM

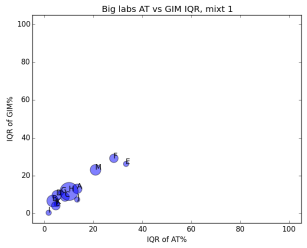


(d) Mixt. 6, AT

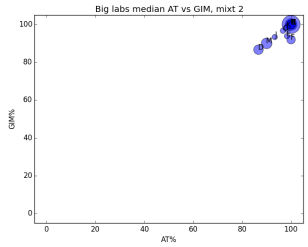
Figure 36: Boxplots for 3-person mixtures $5 - 6$ of the thirteen labs of size five or greater, giving the distributions of each lab's respective examiners' scores. Red lines indicate median scores, boxes delimit the interquartile range, with outliers beyond it. The left column displays the GIM or precision scores from each lab, while the right column displays the AT or accuracy scores.



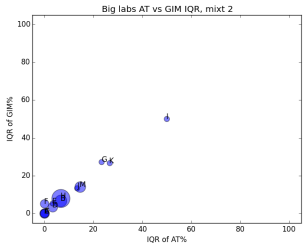
(a) Mixt. 1, median GIM vs AT



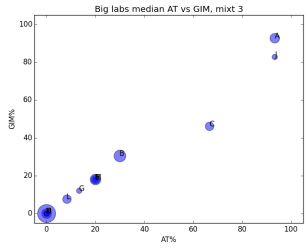
(b) Mixt. 1, IQR GIM vs AT



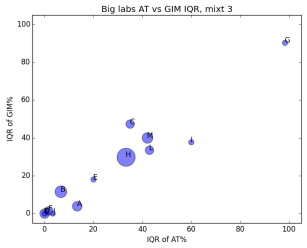
(c) Mixt. 2, median GIM vs AT



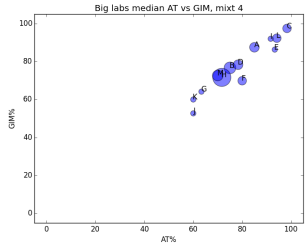
(d) Mixt. 2, IQR GIM vs AT



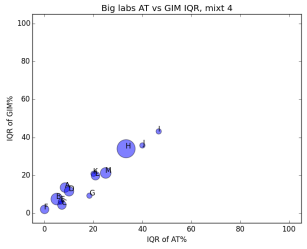
(e) Mixt. 3, median GIM vs AT



(f) Mixt. 3, IQR GIM vs AT

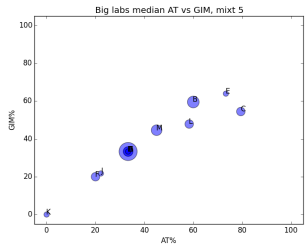


(g) Mixt. 4, median GIM vs AT

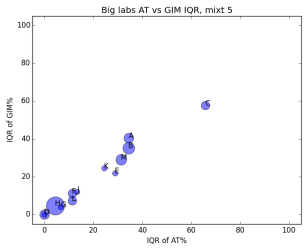


(h) Mixt. 4, IQR GIM vs AT

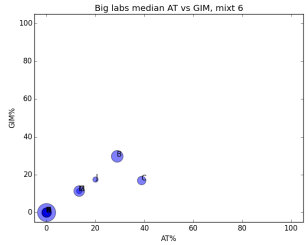
Figure 37: Scatterplots for 2-person mixtures $1-4$ of the thirteen labs of size five or greater, giving the performance of GIM vs AT scores. The radius of each dot is proportional to the number of examiners in the lab. The left column displays the median scores from each lab, while the right column displays the IQR scores.



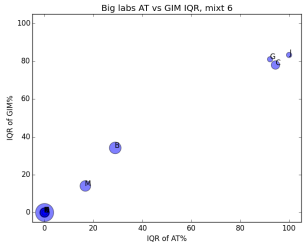
(a) Mixt. 5, median GIM vs AT



(b) Mixt. 5, IQR GIM vs AT



(c) Mixt. 6, median GIM vs AT



(d) Mixt. 6, IQR GIM vs AT

Figure 38: Scatterplots for 3-person mixtures $5-6$ of the thirteen labs of size five or greater, giving the performance of GIM vs AT scores. The radius of each dot is proportional to the number of examiners in the lab. The left column displays the median scores from each lab, while the right column displays the IQR scores.

Fig. 37 shows the wide scatter of lab scores, with median and IQR scores ranging across the spectrum from near 0% to near 100%.

The median scores for mixture 4, a 3:1 mixture, improves to levels similar to to mixture 1, a 3.5:1 mixture. Instead of a spread encompassing the entire spectrum found in mixture 3, the median and IQR scores span about half the spectrum, from 50% to 100% for median scores to 0% to 50% for IQR scores, as seen in Fig. 37. The points falling close to the identity diagonal in Fig. 37 again points to the correlation in GIM and AT scores.

The lowered scores of Mixtures 5 and 6 in Fig. 38 reflect the increased complexity of a 3-person mixture. Results indicate that most labs do not attempt to deconvolute such mixtures. Even with a provided major profile for mixture 5, a majority of the labs had a median GIM at or below 33%, the score of just reporting back the given profile without any further deconvolution. A few labs deconvolute the minor profiles, with the GIM score in general lagging slightly behind the AT score. This indicates that the minority of labs that deconvolute mixture 5 are more cautious, reporting more possible combinations than

in other mixtures.

Mixture 6, as a 1:1:1 mixture with no reference given, has unsurprisingly the lowest scores, with a majority of the labs reporting "Inconclusive" for the entire mixture. A few labs, however, have a majority of examiners deconvoluting the mixture with varying degrees of success, as seen by a median GIM score above 0 in Fig. 36. Within these labs are a few examiners achieving both a relatively high GIM and AT score.

6.7 Discussion

6.7.1 First goal: uncovering the absolute state of mixture interpretation

The results from visualizing the boxplots for the six mixtures shed light on the absolute state of DNA mixture interpretation (the first major goal of this study). Having a reference profile has a marked positive effect on interpretability, increasing both GIM and AT scores such that the two-person Mixture 2 has by far the best results with respect to both median and IQR scores, and the three-person Mixture 5 has similar scores to hardest two-person mixture. Unsurprisingly, peak height, cited in the survey as the most influential factor in interpretation, also plays a significant role in quality of interpretation. Mixtures 1 and 4 having much better results than mixture 3, whose average RFU per allele at 309 is just barely over the stochastic threshold of 300.

Results generally indicate that the two-person mixtures given were interpretable by a sizable number of examiners; all four mixtures had at least 25% of examiners receiving an AT score of at least 0.8. Since a full deconvolution of the two-person mixtures seems within the realm of possibility, efforts to improve the state of DNA mixture interpretation could focus on training examiners to confidently handle similar two-person mixtures. While the majority do very well in Mixtures 1, 2 and 4 (with median GIM and AT scores over 0.8), a majority of examiners seem to falter with mixture 3 (median score of around 0.3). Thus, particular emphasis in protocols and training may need to be put concerning mixtures of lower peak height, so that examiners across the board are equipped to deconvolute mixtures near but still above stochastic threshold.

The consensus on three-person mixtures, however, is that they are generally untouchable,

especially without a reference and/or with lower peak heights. The majority of examiners for Mixture 5 merely report back the reference provided without further attempt to deconvolute the rest of the mixture. With Mixture 6, the majority do not touch it at all, with a median GIM score of 0 indicating that at least half of all examiners mark all loci as 'Inconclusive'.

Although a majority of the examiners do not attempt to deconvolute either Mixtures 5 or 6, both mixtures nevertheless have a few examiners that succeed in doing so. This indicates that successful deconvolution of three-person mixtures, even with Mixture 6's low peak heights and lack of a clear major profile, is not impossible. However, whether these high scores represent the state of the art, or are going beyond the limits of responsible analysis given the numerous known mixture pitfalls that exist, is an open question. If the former, whether or not the techniques of these successful examiners can be codified and reproduced in other examiners is also an open question.

6.7.2 Second goal: uncovering the relative state of each lab's mixture interpretation

The second major goal of this study is to identify the states of individual labs. Coupling this with the first goal of identifying the overall state of DNA mixture interpretation enables pinpointing areas of improvement.

We found that overall, among the larger labs, a strong correlation (Spearman's coefficient > 0.9) exists between the median GIM and AT score. Thus, although discussions regarding accuracy and precision often involve a tradeoff between the two qualities, our Results indicate in the case of DNA mixture interpretation, this is not so. Some labs at one end of the spectrum are able to achieve high levels of both accuracy and precision, whereas others have relatively low levels of both. This surprising trend seems to indicate strongly that some labs have superior training and protocols, and are to be emulated. Because these are median lab scores, they are not outliers or flukes of one or two high-scoring examiners, but represent the state of an entire lab of at least five examiners. We may consider these labs that are consistently able to deconvolute mixtures with high precision and accuracy as the reproducible standard that other labs can be brought up to. Hence, goals for each lab should include improving median scores to at least the level of these labs.

Another mark of superior in-house training may be seen in labs' IQR numbers for both GIM and AT scores, with a smaller range potentially indicating stronger training and/or clearer protocols. Since we found no (strong) correlation between scores and years of experience (Spearman's coefficient), we posit the range of scores found within a lab are not due to differing levels of experience but to other factors such as in-house training and quality control. Therefore, another laboratory goal should be to decrease the range of scores from each constituent examiner.

Furthermore, labs often have a tight IQR in one mixture, but a much larger one in another, with a different lab experiencing the exact opposite. Hence, the areas of consensus (and the lack thereof) differ by labs, and potentially point to specific areas of ambiguities or difficulties unique to each lab. More in depth analysis is planned, with the results made available to each lab and generalized trends presented to the forensic community.

6.8 Conclusion

Forensic DNA analysis has come a long way since its start in the mid 1980's, with advancing techniques allowing ever more detailed and sensitive analysis. Parallel to its development has been its growing importance to criminal justice in the identifying or excluding of individuals. Since the results of DNA interpretation often have profound and long-lasting repercussions, that interpretation should be as objective, reproducible and error-free as possible.

Like all other fields, however, the actual state of the art for DNA interpretation lags behind its ideal, with known errors and biases present. This is especially the case when the DNA being interpreted is sourced from more than one individual and collected as a mixture. The complexity of DNA mixtures carries its own unique complications and potential sources of error.

This study, in inviting interpretations from over 180 examiners from over 55 laboratories of the .fsa files of six carefully curated DNA mixtures, achieves two major goals. First, it illuminates the overall state of DNA mixture interpretation, in order to better understand the current limits of analysis. Namely, it highlights the importance of a reference profile and of strong peak heights in the interpretability of a mixture. It also suggests that two-person

mixtures with signal peaks above stochastic threshold are generally interpretable, while three-person mixtures are currently beyond the scope of most examiners. Results from this study also indicate for all but the easiest and hardest mixtures, a significant amount of variability exists both within and between laboratories.

Second, it provides a detailed summary of the state of individual labs. Taken with respect to the first goal, the second implies ways for labs to improve both their overall scores (median) and their consistency in achieving them (IQR). (More in-depth analysis to identify sources of errors, be they random or systematic, lab-specific or systemwide, is currently underway.) In the this way, the bar set by the best-practice labs can become the standard in the general community, and thus the DNA forensic community as a whole can advance in measurable steps.

Also of potential value is the continuation of controlled studies like this or the continued availability of test mixtures for ongoing evaluation and benchmarking. While laboratories can conduct their own in-house quality control, they will be able to decrease their IQR score by addressing variability in examiner interpretations. The need to improve their overall median score, however, may not be apparent without the comparison to other forensic laboratories that are fully deconvoluting the identical mixtures successfully. The ability to do so can only be helped by the sharing of best practices and techniques within the DNA community [154, 51, 125]. Thus, there is a need for resources and feedback that reach beyond the scope of individual labs, such that successful methods in one lab can become prevalent methods in the general community.

APPENDIX A

CHP. 1 SUPPLEMENTARY INFORMATION

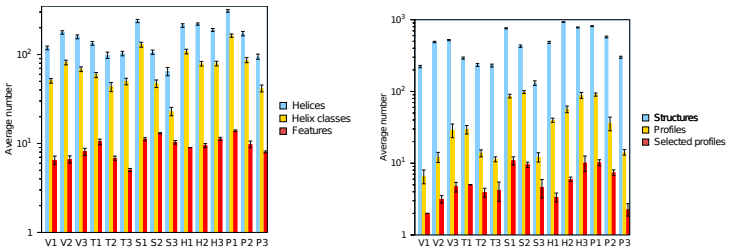


Figure 39: The average number of (left) helices, helix classes and features across 25 samples, and (right) structures, profiles and selected profiles across 25 samples, with bars indicating standard deviation. Log scale is used for additional clarity.

Sequence	Helices		Helix Classes		Features		H:HC		HC:F		H:F	
	Ave	Stdev	Ave	Stdev	Ave	Stdev	Ave	Stdev	Ave	Stdev	Ave	Stdev
V1	118.8	5.7	50.7	3.1	6.5	0.7	2.4	0.2	7.9	1.2	18.5	2.7
V2	177.1	7.9	81.5	4.9	6.6	0.6	**2.2**	0.1	12.4	1.4	**26.9**	2.6
V3	135.1	7.0	**53.1**	4.3	8.0	0.0	2.6	0.2	**6.6**	0.5	**16.9**	0.9
V4	158.3	8.1	68.5	4.2	**8.1**	0.7	2.3	0.1	8.6	1.1	19.8	2.2
V5	164.8	6.9	84.5	3.9	7.4	0.5	2.0	0.1	11.5	0.9	22.5	1.8
T1	**133.7**	6.5	59.0	3.9	10.4	0.8	2.3	0.1	5.7	0.6	12.9	1.4
T2	88.2	8.1	45.6	4.7	8.1	0.6	1.9	0.1	5.7	0.7	10.9	1.3
T3	98.4	8.0	43.4	5.1	6.8	0.4	2.3	0.1	6.3	0.7	14.4	1.2
T4	**48.4**	4.6	24.2	2.6	7.0	0.2	2.0	0.1	3.4	0.4	6.9	0.6
T5	102.7	6.0	49.8	4.3	**5.0**	0.2	2.1	0.1	9.9	0.9	20.4	1.3
S1	238.4	10.1	129.2	8.2	11.2	0.5	1.8	0.1	11.6	1.0	21.4	1.4
S2	**321.3**	10.7	**178.4**	8.3	12.8	0.7	**1.8**	0.1	**14.0**	1.1	25.3	1.7
S3	106.4	5.9	47.2	4.6	13.0	0.2	2.3	0.2	3.6	0.4	8.2	0.5
S4	138.3	8.7	68.3	5.9	**16.0**	0.2	2.0	0.1	4.3	0.4	8.7	0.6
S5	64.6	7.0	**23.0**	2.5	10.3	0.5	**2.8**	0.2	**2.2**	0.3	**6.3**	0.7

Table 16: Data for Fig. 6a: the average and standard deviation in number of helices, helix classes and features, with amplification ratios calculated as average number of helices to helix classes, helix classes to features, and helices to features. Median, minimum and maximum values for averages are bolded.

Sequence	Structures		Profiles		S. Profiles		S:P		P:SP		S:SP	
	Ave	Stdev	Ave	Stdev	Ave	Stdev	Ave	Stdev	Ave	Stdev	Ave	Stdev
V1	222.5	8.3	**6.6**	1.4	**2.0**	0.0	35.9	10.3	3.3	0.7	111.2	4.1
V2	489.8	10.9	12.2	2.0	3.2	0.4	**41.2**	6.7	3.9	0.5	**156.7**	15.0
V3	522.7	13.6	23.2	1.6	6.2	0.4	22.7	1.7	3.7	0.4	84.1	5.8
V4	519.5	10.1	28.9	6.3	4.7	0.7	18.9	4.4	6.3	1.8	114.1	20.0
V5	457.1	13.3	19.6	2.0	3.4	0.6	23.6	2.3	5.9	0.8	139.1	19.0
T1	292.1	9.8	29.5	4.0	5.0	0.0	10.1	1.5	**5.9**	0.8	58.4	2.0
T2	174.0	9.0	**22.6**	2.4	2.0	0.2	7.8	1.0	**11.2**	1.5	85.8	6.9
T3	234.7	11.1	13.8	1.5	3.9	0.6	17.2	2.0	3.6	0.7	61.0	8.6
T4	138.0	7.6	13.2	1.7	3.0	0.0	**10.6**	1.3	4.4	0.6	46.0	2.5
T5	229.7	10.8	11.4	0.9	4.2	1.3	20.3	1.7	3.1	1.3	61.8	24.8
S1	759.1	12.9	86.4	5.1	10.8	1.4	8.8	0.5	8.1	1.1	**71.3**	9.8
S2	**899.3**	7.9	88.2	15.2	**12.8**	1.6	10.5	1.8	7.0	1.8	71.4	9.2
S3	**427.9**	16.1	98.2	4.3	9.5	0.8	**4.4**	0.2	10.4	1.1	45.3	4.8
S4	508.4	14.5	**113.6**	6.9	12.2	2.0	4.5	0.3	9.6	1.6	42.9	7.6
S5	**131.0**	9.6	12.2	1.8	**4.6**	1.3	11.0	1.8	**3.0**	1.3	**32.0**	13.8

Table 17: Data for Fig. 6b: the average and standard deviation in the number of collections of structures, profiles and selected profiles, with amplification ratio calculated as average number of structures to profiles, profiles to selected profiles, and structures to selected profiles. Median, minimum and maximum values for averages are bolded.

Sequence	By Features		By Selected Profiles	
	Ave	Stdev	Ave	Stdev
V1	95.8	0.006	94.2	0.015
V2	90.3	0.010	89.5	0.019
V3	94.9	0.003	81.1	0.019
V4	89.5	0.010	80.8	0.055
V5	92.3	0.005	78.4	0.020
T1	83.7	0.029	82.6	0.027
T2	94.6	0.008	**81.9**	0.017
T3	94.8	0.006	92.9	0.020
T4	98.5	0.002	88.9	0.010
T5	92.8	0.005	97.3	0.017
S1	82.7	0.011	65.0	0.037
S2	**72.8**	0.013	**62.7**	0.081
S3	95.4	0.003	73.5	0.021
S4	**94.0**	0.006	71.2	0.042
S5	**99.4**	0.001	**97.0**	0.022

Table 18: Average (with standard deviation) percent coverage across 25 runs of helices by features, and structures by selected profiles. Median, minimum and maximum values for averages are bolded.

Sequence	Features		Selected profiles	
	Ave	Stdev	Ave	Stdev
V1	**0.938**	0.032	**1.000**	0.000
V2	0.944	0.055	0.809	0.066
V3	**1.000**	0.000	0.974	0.046
V4	0.952	0.038	0.895	0.075
V5	0.971	0.038	0.934	0.100
T1	0.957	0.011	0.805	0.098
T2	**0.981**	0.050	0.987	0.063
T3	**0.981**	0.008	0.854	0.081
T4	0.995	0.024	**1.000**	0.000
T5	0.994	0.031	0.839	0.098
S1	0.983	0.027	0.827	0.062
S2	0.959	0.028	**0.747**	0.066
S3	0.997	0.013	0.958	0.027
S4	0.998	0.000	**0.881**	0.061
S5	0.980	0.029	0.789	0.088

Table 19: Average reproducibility across 25 runs with standard deviation. Median, minimum and maximum values for averages are bolded.

Helix class	V1	V2	V3	V4	V5
1	(1,16,6)	(1,25,8)	(1,23,8)	(59,88,10)	(2,23,7)
2	(38,56,7)	(77,102,10)	(45,62,7)	(2,14,5)	(75,101,9)
3	(64,88,10)	(47,64,7)	(73,99,11)	(37,53,7)	(46,64,7)
4	(31,63,6)	(32,43,4)	(30,41,4)	(22,33,4)	(31,42,4)
5	(19,28,3)	(27,47,5)	(25,45,5)	(54,93,4)	(84,93,2)
6	(21,30,1)	(51,75,7)	(25,70,3)	(1,16,1)	(26,46,5)
7	(23,28,1)	(29,71,3)	(5,20,5)	(70,77,1)	(24,68,2)
8	(23,34,4)	(48,77,3)	(69,103,3)	(59,89,3)	(29,68,2)
9				(17,37,3)	

Table 20: Potential features for Qrr RNA sequences found across 25 runs, in (i,j,k) notation for associated maximal helix.

Helix class	T1	T2	T3	T4	T5
1	(1,13,4)	(1,73,7)	(1,71,7)	(1,73,8)	(1,72,7)
2	(15,70,6)	(33,44,4)	(23,47,3)	(38,56,6)	(49,65,5)
3	(48,64,5)	(30,48,3)	(27,43,5)	(29,65,4)	(26,44,6)
4	(36,47,4)	(7,31,7)	(48,64,5)	(10,26,4)	(9,26,5)
5	(23,34,3)	(50,66,5)	(7,22,6)	(34,60,3)	(31,40,3)
6	(26,42,5)	(15,24,3)	(31,40,3)	(15,21,2)	(52,61,3)
7	(29,54,5)	(29,50,2)	(19,51,3)	(26,68,2)	
8	(25,59,6)	(7,66,4)		(35,60,3)	
9	(21,47,2)	(52,63,4)			
10	(24,44,1)	(26,52,3)			
11	(14,72,5)				

Table 21: Potential features for tRNA sequences found across 25 runs, in (i,j,k) notation for associated maximal helix.

Helix class	S1	S2	S3	S4	S5
1	(1,119,10)	(1,118,9)	(1,118,9)	(84,93,3)	(3,133,9)
2	(33,88,10)	(85,92,2)	(25,63,7)	(1,123,2)	(12,123,6)
3	(95,104,3)	(78,98,6)	(66,109,7)	(66,109,7)	(22,73,8)
4	(20,31,4)	(67,108,4)	(84,93,3)	(16,62,6)	(75,117,10)
5	(61,77,5)	(16,62,5)	(78,99,5)	(1,118,9)	(86,106,8)
6	(91,107,3)	(14,65,4)	(31,44,4)	(14,65,2)	(37,57,4)
7	(48,57,3)	(10,110,3)	(22,53,3)	(25,52,2)	(43,51,3)
8	(89,108,4)	(72,104,1)	(28,48,2)	(29,48,4)	(32,62,3)
9	(46,55,3)	(41,59,5)	(14,65,2)	(78,99,5)	(93,100,2)
10	(10,22,5)	(51,66,6)	(76,100,1)	(31,44,4)	(29,67,3)
11	(60,69,3)	(30,36,2)	(75,102,2)	(35,42,2)	(35,61,2)
12	(25,108,4)	(42,74,5)	(29,48,4)	(76,100,1)	
13		(72,103,2)	(73,103,2)	(81,95,2)	
14		(31,39,3)	(35,42,2)	(78,98,2)	
15		(25,31,2)		(75,102,2)	
16				(73,103,2)	

Table 22: Potential features for 5S RNA sequences found across 25 runs, in (i,j,k) notation for associated maximal helix.

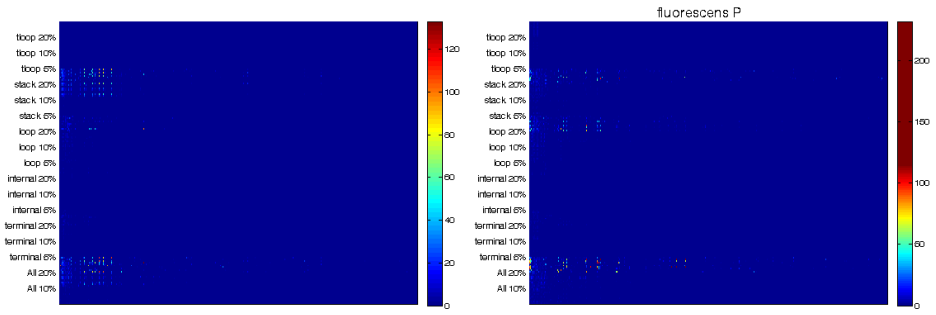
APPENDIX B

PARAMETER SUBSET DATA

Another question we probe is whether all NNTM parameters are equally influential, or if results are more sensitive to perturbation of select parameters. Because the actual number of parameters number in the thousands, some of which may be only accessed in rare cases, we test subsets of parameters that are structurally related, guided roughly by the groupings done by the separate files. Namely, we group the NNTM parameter files into five subcategories: internal (internal loop parameters, composed of the files int11.dat, int21.dat and int22.dat), stack (parameters scoring a helical stack found in stack.dat), loop (loop initiation penalties found in loop.dat), tloop (tetraloop special cases in tloop.dat) and terminal (parameters scoring terminal penalties of loops found in dangle.dat, tstackh.dat and tstacki.dat).

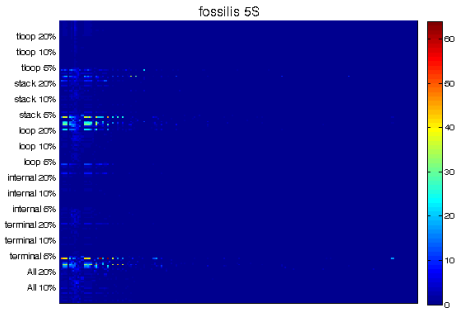
We ran Boltzmann sampling and subsequent profiling on the sequences with the same 5%, 10% and 20% perturbed parameters as before, but only using the perturbed files of interest. All other files not being investigated were kept the same as the original parameter files.

To investigate nuances of parameter influence, we not only calculate total disturbance and κ for the results, but also track the effect on individual helix classes. Figures 40 show heatmaps of all the helix classes ordered in descending frequency, with colors reflecting absolute value changes to $PS(h)$ under a given set of perturbed parameters.

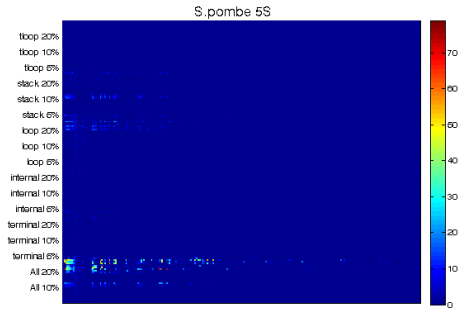


(a) RNAseP A.ferrooxidans X165580

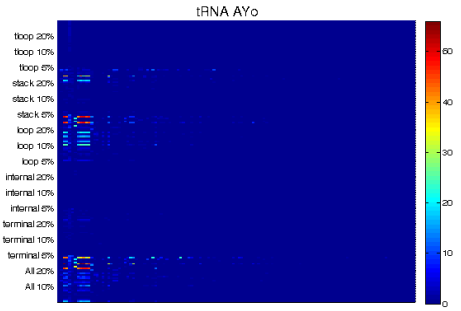
(b) RNAseP P.fluorescens



(c) 5S rRNA M.fossilis V00647



(d) 5S rRNA S.pombe K00570



(e) tRNA AYo AY934018

Figure 40: Heat maps indicating level of output change with perturbation of different subsets of NNTM parameters. Ten different random parameter sets were generated at each perturbation level. Rows start at the bottom; e.g. values inbetween two vertical labels are for the lower of the two labels. The perturb level 'All 5%' is missing from the very bottom of the y-axis, and whose values are reflected in the bottommost 10 rows. The x-axis represents all the original helix classes of the sequence; the color on the y-axis represents the degree of change of that helix class when sampled again under perturbation. The degree of change is in the same units as discussed in the *Biophys. Jour.* paper (Chapter 4). As seen, the subsets with the most effect are the loop and stack parameters.

REFERENCES

[1] Allali, J. and Sagot, M.-F., "A new distance for high level RNA secondary structure comparison," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 2, no. 1, pp. 3–14, 2005.

[2] Anderson, J. W., Haas, P. A., Mathieson, L.-A., Volynkin, V., Lyngsø, R., Tataru, P., and Hein, J., "Oxfold: kinetic folding of RNA using stochastic context-free grammars and evolutionary information," *Bioinformatics*, vol. 29, no. 6, pp. 704–710, 2013.

[3] Asai, K. and Hamada, M., "RNA structural alignments, part II: non-Sankoff approaches for structural alignments," *Meth Mol Biol*, vol. 1097.

[4] Bafna, V., Tang, H., and Zhang, S., "Consensus folding of unaligned RNA sequences revisited," *Jour of Comp Biol*, vol. 13, no. 2, pp. 283–295, 2006.

[5] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., and Nielsen, H., "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinfor*, vol. 16, no. 5, pp. 412–424, 2000.

[6] Balding, D. J. and Buckleton, J., "Interpreting low template DNA profiles," *Forensic Science International: Genetics*, vol. 4, no. 1, pp. 1–10, 2009.

[7] Bardill, J. P. and Hammer, B., "Non-coding sRNAs regulate virulence in the bacterial pathogen Vibrio cholerae," *RNA Biol*, vol. 9, no. 4, pp. 392–401, 2012.

[8] Barrick, J. E., Corbino, K. A., Winkler, W. C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I., and others, "New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control," *Proc Natl Acad Sci USA*, vol. 101, no. 17, pp. 6421–6426, 2004.

[9] Benschop, C. C., Haned, H., de Blaeij, T. J., Meulenbroek, A. J., and Sijen, T., "Assessment of mock cases involving complex low template DNA mixtures: a descriptive study," *Forensic Science International: Genetics*, vol. 6, no. 6, pp. 697–707, 2012.

[10] Bing, T., Yang, X., Mei, H., Cao, Z., and Shangguan, D., "Conservative secondary structure motif of streptavidin-binding aptamers generated by different laboratories," *Bioorg Med Chem*, vol. 18, no. 5, pp. 1798–1805, 2010.

[11] Breaker, R. R., "Are engineered proteins getting competition from RNA?," *Current Opinion in Biotechnology*, vol. 7, no. 4, pp. 442–448, 1996.

[12] Bright, J.-A., Curran, J. M., and Buckleton, J. S., "Investigation into the performance of different models for predicting stutter," *Forensic Science International: Genetics*, vol. 7, no. 4, pp. 422–427, 2013.

[13] Bright, J.-A., Taylor, D., Curran, J. M., and Buckleton, J. S., "Developing allelic and stutter peak height models for a continuous method of DNA interpretation," *Forensic Science International: Genetics*, vol. 7, no. 2, pp. 296–304, 2013.

[14] Bright, J.-A., Taylor, D., Gittelson, S., and Buckleton, J., "The paradigm shift in DNA profile interpretation," *Forensic Science International: Genetics*, vol. 31, pp. e24–e32, 2017.

[15] Brown, J. W., "The ribonuclease P database," *Nucleic Acids Res*, vol. 27, no. 1, pp. 314–314, 1999.

[16] Buckleton, J. S., Curran, J. M., and Gill, P., "Towards understanding the effect of uncertainty in the number of contributors to DNA stains," *Forensic Science International: Genetics*, vol. 1, no. 1, pp. 20–28, 2007.

[17] Budowle, B., Bottrell, M. C., Bunch, S. G., Fram, R., Harrison, D., Meagher, S., Oien, C. T., Peterson, P. E., Seiger, D. P., Smith, M. B., and others, "A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement," *Journal of Forensic Sciences*, vol. 54, no. 4, pp. 798–809, 2009.

[18] Budowle, B., Onorato, A. J., Callaghan, T. F., Manna, A. D., Gross, A. M., Guerrieri, R. A., Luttman, J. C., and McClure, D. L., "Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework," *Journal of Forensic Sciences*, vol. 54, no. 4, pp. 810–821, 2009.

[19] Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P., and Bateman, A., "Rfam 11.0: 10 years of RNA families," *Nuc Acids Res*, vol. 41, no. D1, pp. D226–D232, 2012.

[20] Caliński, T. and Harabasz, J., "A dendrite method for cluster analysis," *Comm Statistics-theory Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[21] Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M., Du, Y., Feng, B., Lin, N., Madabusi, L. V., Müller, K. M., and others, "The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs," *BMC bioinformatics*, vol. 3, no. 1, p. 2, 2002.

[22] Chan, C. Y., Lawrence, C. E., and Ding, Y., "Structure clustering features on the Sfold Web server.," *Bioinformatics*, vol. 21, no. 20, pp. 3926–3928, 2005.

[23] Clayton, T., Whitaker, J., Sparkes, R., and Gill, P., "Analysis and interpretation of mixed forensic stains using DNA STR profiling," *Forensic Science International*, vol. 91, no. 1, pp. 55–70, 1998.

[24] Couzin, J., "Small RNAs make big splash," *Science*, vol. 298, no. 5602, pp. 2296–2297, 2002.

[25] Datta, S. and Datta, S., "Comparisons and validation of statistical clustering techniques for microarray gene expression data," *Bioinformatics*, vol. 19, no. 4, pp. 459–466, 2003.

[26] Davis, J. H. and Szostak, J. W., "Isolation of high-affinity GTP aptamers from partially structured RNA libraries," *Proc Natl Acad Sci USA*, vol. 99, no. 18, pp. 11616–11621, 2002.

[27] Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M., "Accurate SHAPE-directed RNA structure determination," *Proc Natl Acad Sci*, vol. 106, no. 1, pp. 97–102, 2009.

[28] Del Campo, C., Bartholomäus, A., Fedyunin, I., and Ignatova, Z., "Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function," *PLoS Genet*, vol. 11, no. 10, p. e1005613, 2015.

[29] Ding, Y., Chan, C. Y., and Lawrence, C. E., "Sfold web server for statistical folding and rational design of nucleic acids," *Nucleic Acids Res*, vol. 32, no. suppl 2, pp. W135–W141, 2004.

[30] Ding, Y., Chan, C. Y., and Lawrence, C. E., "Rna secondary structure prediction by centroids in a Boltzmann weighted ensemble," *RNA*, vol. 11, no. 8, pp. 1157–1166, 2005.

[31] Ding, Y., Chan, C. Y., and Lawrence, C. E., "Clustering of RNA secondary structures with application to messenger RNAs," *J Mol Biol*, vol. 359, no. 3, pp. 554 –71, 2006.

[32] Ding, Y. and Lawrence, C. E., "A statistical sampling algorithm for RNA secondary structure prediction," *Nucleic Acids Res*, vol. 31, no. 24, pp. 7280–7301, 2003.

[33] Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., and Assmann, S. M., "In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features," *Nature*, vol. 505, no. 7485, pp. 696–700, 2014.

[34] Do, C. B., Woods, D. A., and Batzoglou, S., "CONTRAfold: RNA secondary structure prediction without physics-based models," *Bioinformatics*, vol. 22, no. 14, pp. e90–e98, 2006.

[35] Doshi, K. J., Cannone, J. J., Cobaugh, C. W., and Gutell, R. R., "Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction," *BMC bioinformatics*, vol. 5, no. 1, p. 105, 2004.

[36] Doudna, J. A., "Structural genomics of RNA," *Nat Struct Biol*, vol. 7, no. 11, pp. 954–956, 2000.

[37] Dror, I. E., "Cognitive forensics and experimental research about bias in forensic casework," *Science and Justice*, vol. 52, no. 2, pp. 128–130, 2012.

[38] Dror, I. E. and Hampikian, G., "Subjectivity and bias in forensic DNA mixture interpretation," *Science and Justice*, vol. 51, no. 4, pp. 204–208, 2011.

[39] Duewer, D. L., Kline, M. C., Redman, J. W., and Butler, J. M., "Nist mixed stain study 3: signal intensity balance in commercial short tandem repeat multiplexes," *Analytical chemistry*, vol. 76, no. 23, pp. 6928–6934, 2004.

[40] Duewer, D. L., Kline, M. C., Redman, J. W., Newall, P. J., and Reeder, D., "NIST mixed stain studies# 1 and# 2: interlaboratory comparison of DNA quantification practice and short tandem repeat multiplex performance with multiple-source samples," *Journal of Forensic Science*, vol. 46, no. 5, pp. 1199–1210, 2001.

[41] Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G., *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. New York: Cambridge University Press, 1998.

[42] Edgar, R. C., "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res*, vol. 32, no. 5, pp. 1792–1797, 2004.

[43] Ester, M., Kriegel, H., Sander, J., Xu, X., and others, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *KDD*, vol. 96, pp. 226–231, 1996.

[44] Gan, H. H., Fera, D., Zorn, J., Shiffeldrim, N., Tang, M., Laserson, U., Kim, N., and Schlick, T., "RAG: RNA-As-Graphs databaseconcepts, analysis, and features," *Bioinformatics*, vol. 20, no. 8, pp. 1285–1291, 2004.

[45] Gardner, P. P., Wilm, A., and Washietl, S., "A benchmark of multiple sequence alignment programs upon structural RNAs," *Nucleic Acids Res*, vol. 33, no. 8, pp. 2433–2439, 2005.

[46] Gardner, P. and Giegerich, R., "A comprehensive comparison of comparative RNA structure prediction approaches," *BMC Bioinformatics*, vol. 5, no. 1, p. 140, 2004.

[47] Ge, P. and Zhang, S., "Computational analysis of RNA structures with chemical probing data," *Methods*, vol. 79, pp. 60–66, 2015.

[48] Gevertz, J., Gan, H. H., and Schlick, T., "In vitro RNA random pools are not structurally diverse: a computational analysis," *RNA*, vol. 11, no. 6, pp. 853–863, 2005.

[49] Giegerich, R., Voss, B., and Rehmsmeier, M., "Abstract shapes of RNA," *Nucleic Acids Res*, vol. 32, no. 16, pp. 4843–4851, 2004.

[50] Giegerich, R., Voss, B., and Rehmsmeier, M., "Abstract shapes of RNA," *Nucleic Acids Res*, vol. 32, no. 16, pp. 4843–4851, 2004.

[51] Gill, P., Brenner, C. H., Buckleton, J. S., Carracedo, A., Krawczak, M., Mayr, W., Morling, N., Prinz, M., Schneider, P. M., and Weir, B., "DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures," *Forensic science international*, vol. 160, no. 2-3, pp. 90–101, 2006.

[52] Gill, P., Whitaker, J., Flaxman, C., Brown, N., and Buckleton, J., "An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA," *Forensic Science International*, vol. 112, no. 1, pp. 17–40, 2000.

[53] Gorodkin, J., Stricklin, S. L., and Stormo, G. D., "Discovering common stem–loop motifs in unaligned RNA sequences," *Nucleic Acids Res*, vol. 29, no. 10, pp. 2135–2144, 2001.

[54] Gratton, S., "On the condition number of linear least squares problems in a weighted Frobenius norm," *BIT Numerical Mathematics*, vol. 36, no. 3, pp. 523–530, 1996.

[55] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R., "Rfam: an RNA family database," *Nucleic Acids Res*, vol. 31, no. 1, pp. 439–441, 2003.

[56] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R., "Rfam: an RNA family database," *Nucleic Acids Res*, vol. 31, no. 1, pp. 439–441, 2003.

[57] Hamada, M., Tsuda, K., Kudo, T., Kin, T., and Asai, K., "Mining frequent stem patterns from unaligned RNA sequences," *Bioinfor*, vol. 22, no. 20, pp. 2480–2487, 2006.

[58] Hammer, B. K. and Bassler, B. L., "Regulatory small RNAs circumvent the conventional quorum sensing pathway in pandemic Vibrio cholerae," *Proc Natl Acad Sci*, vol. 104, no. 27, pp. 11145–11149, 2007.

[59] Havgaard, J. H., Lyngsø, R. B., Stormo, G. D., and Gorodkin, J., "Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%," *Bioinfor*, vol. 21, no. 9, pp. 1815–1824, 2005.

[60] Havgaard, J. H. and Gorodkin, J., "RNA structural alignments, part I: Sankoff-based approaches for structural alignments," *Meth Mol Biol*, vol. 1097.

[61] Higham, D. J., "Condition numbers and their condition numbers," *Linear Alg App*, vol. 214, pp. 193–213, 1995.

[62] Hochsmann, M., Toller, T., Giegerich, R., and Kurtz, S., "Local similarity in RNA secondary structures," in *Bioinfor Conf, 2003. CSB 2003. Proc of the 2003 IEEE*, pp. 159–168, IEEE, 2003.

[63] Hofacker, I. L., Bernhart, S. H., and Stadler, P. F., "Alignment of RNA base pairing probability matrices," *Bioinfor*, vol. 20, no. 14, pp. 2222–2227, 2004.

[64] Hofacker, I. L., "Vienna RNA secondary structure server," *Nucleic Acids Res*, vol. 31, no. 13, pp. 3429–3431, 2003.

[65] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P., "Fast folding and comparison of RNA secondary structures," *Monatshefte für Chemie/Chemical Monthly*, vol. 125, no. 2, pp. 167–188, 1994.

[66] Hoinka, J., Zotenko, E., Friedman, A., Sauna, Z. E., and Przytycka, T. M., "Identification of sequence–structure RNA binding motifs for SELEX-derived aptamers," *Bioinformatics*, vol. 28, no. 12, pp. i215–i223, 2012.

[67] Houck, M. M., *Professional issues in forensic science*. Academic Press, 2015.

[68] Huang, J., Backofen, R., and Voss, B., "Abstract folding space analysis based on helices," *RNA*, vol. 18, no. 12, pp. 2135–2147, 2012.

[69] Huang, J. and Voss, B., "Analysing RNA-kinetics based on folding space abstraction," *BMC bioinformatics*, vol. 15, no. 1, p. 60, 2014.

[70] Huizenga, D. E. and Szostak, J. W., "A DNA aptamer that binds adenosine and ATP," *Biochemistry*, vol. 34, no. 2, pp. 656–665, 1995.

[71] Huynen, M., Gutell, R., and Konings, D., "Assessing the reliability of RNA folding using statistical mechanics1," *J of Mol Biol*, vol. 267, no. 5, pp. 1104–1112, 1997.

[72] Isambert, H., "The jerky and knotty dynamics of RNA," *Methods*, vol. 49, no. 2, pp. 189–196, 2009.

[73] J. Patrick Bardill, X. Z. and Hammer, B. K., "The Vibrio cholerae quorum sensing response is mediated by Hfq-dependent sRNA/mRNA base pairing interactions," *Mol Microbiol*, vol. 80, no. 5, pp. 1381–1394, 2011.

[74] Jacobson, A. B. and Zuker, M., "Structural analysis by energy dot plot of a large mRNA," *J Mol Biol*, vol. 233, no. 2, pp. 261–269, 1993.

[75] Jaeger, J. A., Turner, D. H., and Zuker, M., "Improved predictions of secondary structures for RNA," *Proc Natl Acad Sci USA*, vol. 86, no. 20, pp. 7706–7710, 1989.

[76] Ji, Y., Xu, X., and Stormo, G. D., "A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences," *Bioinfor*, vol. 20, no. 10, pp. 1591–1602, 2004.

[77] Jobling, M. A. and Gill, P., "Encoded evidence: DNA in forensic analysis," *Nature Reviews Genetics*, vol. 5, no. 10, p. 739, 2004.

[78] Joshi, A. and Kaur, R., "A review: Comparative study of various clustering techniques in data mining," *Intl Jour of Advd Res in Comp Sci and Software Eng*, vol. 3, no. 3, 2013.

[79] Kaufman, L. and Rousseeuw, P. J., *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009.

[80] Kaufman, L. and Rousseeuw, P. J., *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009.

[81] Kaye, D. H., "The design of "the first experimental study exploring DNA interpretation"," *Science and Justice*, vol. 52, no. 2, pp. 126–127, 2012.

[82] Keefe, A. D., Pai, S., and Ellington, A., "Aptamers as therapeutics," *Nat Rev Drug Discov*, vol. 9, no. 7, pp. 537–550, 2010.

[83] Kelly, H., Bright, J.-A., Curran, J., and Buckleton, J., "The interpretation of low level DNA mixtures," *Forensic Science International: Genetics*, vol. 6, no. 2, pp. 191–197, 2012.

[84] Kim, N., Fuhr, K. N., and Schlick, T., "Graph applications to RNA structure and function," in *Biophysics of RNA Folding*, pp. 23–51, Springer, 2013.

[85] Kim, N., Petingi, L., and Schlick, T., "Network theory tools for RNA modeling," *WSEAS transactions on mathematics*, vol. 9, no. 12, p. 941, 2013.

[86] Kline, M. C., Duewer, D. L., Redman, J. W., and Butler, J. M., "NIST Mixed Stain Study 3: DNA quantitation accuracy and its influence on short tandem repeat multiplex signal intensity," *Analytical chemistry*, vol. 75, no. 10, pp. 2463–2469, 2003.

[87] Kline, M. C., Duewer, D. L., Redman, J. W., and Butler, J. M., "Results from the NIST 2004 DNA quantitation study," *Journal of Forensic Science*, vol. 50, no. 3, pp. 1–8, 2005.

[88] Kloosterman, A., Sjerps, M., and Quak, A., "Error rates in forensic DNA analysis: definition, numbers, impact and communication," *Forensic science international: Genetics*, vol. 12, pp. 77–85, 2014.

[89] Knudsen, B. and Hein, J., "Pfold: RNA secondary structure prediction using stochastic context-free grammars," *Nucleic Acids Res*, vol. 31, no. 13, pp. 3423–3428, 2003.

[90] Kutchko, K. M., Sanders, W., Ziehr, B., Phillips, G., Solem, A., Halvorsen, M., Weeks, K. M., Moorman, N., and Laederach, A., "Multiple conformations are a conserved and regulatory feature of the RB1 5' UTR," *RNA*, vol. 21, no. 5, pp. 1274–1285, 2015.

[91] Ladd, C., Lee, H. C., Yang, N., and Bieber, F. R., "Interpretation of complex forensic DNA mixtures," *Croatian medical journal*, vol. 42, no. 3, pp. 244–246, 2001.

[92] Lalwani, S., Kumar, R., and Gupta, N., "Sequence-structure alignment techniques for RNA: A comprehensive survey," *Advances in Life Sciences*, vol. 4, no. 1, pp. 21–35, 2014.

[93] Laserson, U., Gan, H. H., and Schlick, T., "Searching for 2D RNA geometries in bacterial genomes," in *Proc 20th Ann Symp Comput Geom*, pp. 373–377, ACM, 2004.

[94] Layton, D. and Bundschuh, R., "A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation," *Nucleic Acids Res*, vol. 33, no. 2, pp. 519–524, 2005.

[95] Le, S.-Y., Chen, J.-H., and Maizel, J. V., "Prediction of alternative RNA secondary structures based on fluctuating thermodynamic parameters," *Nucleic Acids Res*, vol. 21, no. 9, pp. 2173–2178, 1993.

[96] Le, S.-Y., Nussinov, R., and Maizel, J. V., "Tree graphs of RNA secondary structures and their comparisons," *Computers and Biomedical Research*, vol. 22, no. 5, pp. 461–473, 1989.

[97] Leclair, B., Frégeau, C. J., Bowen, K. L., and Fourney, R. M., "Systematic analysis of stutter percentages and allele peak height and peak area ratios at heterozygous STR loci for forensic casework and database samples," *Journal of Forensic Science*, vol. 49, no. 5, pp. JFS2003312–13, 2004.

[98] LeCuyer, K. A. and Crothers, D. M., "The Leptomonas collosoma spliced leader RNA can switch between two alternate structural forms," *Biochemistry*, vol. 32, no. 20, pp. 5301–5311, 1993.

[99] Lenhof, H.-P., Reinert, K., and Vingron, M., "A polyhedral approach to RNA sequence structure alignment," *J of Comp Bio*, vol. 5, no. 3, pp. 517–530, 1998.

[100] Lenz, D. H., Mok, K. C., Lilley, B. N., Kulkarni, R. V., Wingreen, N. S., and Bassler, B. L., "The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae," *Cell*, vol. 117, no. 1, pp. 69–82, 2004.

[101] Lindgreen, S., Gardner, P. P., and Krogh, A., "MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing," *Bioinfor*, vol. 23, no. 24, pp. 3304–3311, 2007.

[102] Liu, J., Wang, J. T., Hu, J., and Tian, B., "A method for aligning RNA secondary structures and its application to RNA motif detection," *BMC bioinfor*, vol. 6, no. 1, p. 89, 2005.

[103] Mandal, M. and Breaker, R. R., "Gene regulation by riboswitches," *Nat Rev Mol Cell Biol*, vol. 5, no. 6, pp. 451–463, 2004.

[104] Mathews, D. H., "Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization," *RNA*, vol. 10, no. 8, pp. 1178–1190, 2004.

[105] Mathews, D. H., "Revolutions in RNA secondary structure prediction," *J Mol Biol*, vol. 359, no. 3, pp. 526–532, 2006.

[106] Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H., "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure," *J Mol Biol*, vol. 288, no. 5, pp. 911–940, 1999.

[107] Mathews, D. H. and Turner, D. H., "Dynalign: an algorithm for finding the secondary structure common to two RNA sequences," *J Mol Biol*, vol. 317, no. 2, pp. 191–203, 2002.

[108] Mathews, D. H. and Turner, D. H., "Prediction of RNA secondary structure by free energy minimization," *Curr Opin Struct Biol*, vol. 16, no. 3, pp. 270–278, 2006.

[109] McCaskill, J. S., "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," *Biopolymers*, vol. 29, no. 6-7, pp. 1105–1119, 1990.

[110] Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and Weeks, K. M., "RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE)," *J Amer Chem Soc*, vol. 127, no. 12, pp. 4223–4231, 2005.

[111] Miller, M. B. and Bassler, B. L., "Quorum sensing in bacteria," *Annu Rev Microbiol*, vol. 55, no. 1, pp. 165–199, 2001.

[112] Miyashiro, T., Wollenberg, M. S., Cao, X., Oehlert, D., and Ruby, E. G., "A single qrr gene is necessary and sufficient for LuxO-mediated regulation in *Vibrio fischeri*.," *Mol Microbiol*, vol. 77, pp. 1556–67, Sep 2010.

[113] Montange, R. K. and Batey, R. T., "Riboswitches: emerging themes in RNA structure and function," *Annu. Rev. Biophys.*, vol. 37, pp. 117–133, 2008.

[114] Moulton, V., Zuker, M., Steel, M., Pointon, R., and Penny, D., "Metrics on RNA secondary structures," *J Comp Biol*, vol. 7, no. 1-2, pp. 277–292, 2000.

[115] Mutreja, A., Kim, D. W., Thomson, N. R., Connor, T. R., Lee, J. H., Kariuki, S., Croucher, N. J., Choi, S. Y., Harris, S. R., Lebens, M., and others, "Evidence for several waves of global transmission in the seventh cholera pandemic," *Nature*, vol. 477, no. 7365, pp. 462–465, 2011.

[116] Nimjee, S. M., Rusconi, C. P., and Sullenger, B. A., "Aptamers: an emerging class of therapeutics," *Annu. Rev. Med.*, vol. 56, pp. 555–583, 2005.

[117] Nussinov, R. and Jacobson, A. B., "Fast algorithm for predicting the secondary structure of single-stranded RNA," *Proc Natl Acad Sci*, vol. 77, no. 11, pp. 6309–6313, 1980.

[118] Nutiu, R. and Li, Y., "In vitro selection of structure-switching signaling aptamers," *Angewandte Chemie*, vol. 117, no. 7, pp. 1085–1089, 2005.

[119] Pan, T. and Sosnick, T. R., "Intermediates and kinetic traps in the folding of a large ribozyme revealed by circular dichroism and UV absorbance spectroscopies and catalytic activity," *Nat Struct Mol Biol*, vol. 4, no. 11, pp. 931–938, 1997.

[120] Paoletti, D. R., Doom, T. E., Krane, C. M., Raymer, M. L., and Krane, D. E., "Empirical analysis of the STR profiles resulting from conceptual mixtures," *Journal of Forensic Science*, vol. 50, no. 6, pp. JFS2004475–6, 2005.

[121] Parisien, M. and Major, F., "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data," *Nature*, vol. 452, no. 7183, pp. 51–55, 2008.

[122] Pascali, V. L. and Merigioli, S., "'stochastic' effects at balanced mixtures: A calibration study," *Forensic Science International: Genetics*, vol. 8, no. 1, pp. 113–125, 2014.

[123] Perlin, M. W., Hornyak, J. M., Sugimoto, G., and Miller, K. W., "Trueallele® genotype identification on DNA mixtures containing up to five unknown contributors," *Journal of forensic sciences*, vol. 60, no. 4, pp. 857–868, 2015.

[124] Perlin, M. W. and Sinelnikov, A., "An information gap in DNA evidence interpretation," *PLOS one*, vol. 4, no. 12, p. e8327, 2009.

[125] Prieto, L., Haned, H., Mosquera, A., Crespillo, M., Alemañ, M., Aler, M., Alvarez, F., Baeza-Richer, C., Dominguez, A., Doutremepuich, C., and others, "Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles," *Forensic Science International: Genetics*, vol. 9, pp. 47–54, 2014.

[126] Puch-Solis, R., Rodgers, L., Mazumder, A., Pope, S., Evett, I., Curran, J., and Balding, D., "Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters," *Forensic Science International: Genetics*, vol. 7, no. 5, pp. 555–563, 2013.

[127] Puton, T., Kozlowski, L. P., Rother, K. M., and Bujnicki, J. M., "CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction," *Nucleic Acids Res*, vol. 41, no. 7, pp. 4307–4323, 2013.

[128] Rakay, C. A., Bregu, J., and Grgicak, C. M., "Maximizing allele detection: effects of analytical threshold and DNA levels on rates of allele and locus drop-out," *Forensic Science International: Genetics*, vol. 6, no. 6, pp. 723–728, 2012.

[129] Rand, S., Schürenkamp, M., Hohoff, C., and Brinkmann, B., "The GEDNAP blind trial concept part ii. Trends and developments," *International journal of legal medicine*, vol. 118, no. 2, pp. 83–89, 2004.

[130] Rand, S., Schürenkamp, M., and Brinkmann, B., "The GEDNAP (German DNA profiling group) blind trial concept," *International journal of legal medicine*, vol. 116, no. 4, pp. 199–206, 2002.

[131] Reeder, J., Höchsmann, M., Rehmsmeier, M., Voss, B., and Giegerich, R., "Beyond Mfold: recent advances in RNA bioinformatics," *J Biotechnol*, vol. 124, no. 1, pp. 41–55, 2006.

[132] Rogers, E. and Heitsch, C., "New insights from cluster analysis methods for RNA secondary structure prediction," *Wiley Inter Reviews: RNA*, vol. 7, no. 3, pp. 278–294, 2016.

[133] Rogers, E., Murrugarra, D., and Heitsch, C., "Conditioning and robustness of RNA Boltzmann sampling under thermodynamic parameter perturbations," *Biophysl Jourl*, vol. 113, no. 2, pp. 321–329, 2017.

[134] Rogers, E. and Heitsch, C. E., "Profiling small RNA reveals multimodal substructural signals in a Boltzmann ensemble," *Nucleic Acids Res*, p. gku959, 2014.

[135] Rutherford, S. T., van Kessel, J. C., Shao, Y., and Bassler, B. L., "AphA and LuxR/HapR reciprocally control quorum sensing in vibrios," *Genes Dev*, vol. 25, no. 4, pp. 397–408, 2011.

[136] Sanjuán, R., Cuevas, J. M., Furió, V., Holmes, E. C., and Moya, A., "Selection for robustness in mutagenized RNA viruses," *PLoS Genet*, vol. 3, no. 6, p. e93, 2007.

[137] SantaLucia, J. and Turner, D. H., "Measuring the thermodynamics of RNA secondary structure formation," *Biopolymers*, vol. 44, no. 3, pp. 309–319, 1997.

[138] Schuster, P. and Stadler, P. F., "Discrete models of biopolymers," *Handbook of computational chemistry and biology*, pp. 187–221, 2004.

[139] Serganov, A. and Patel, D. J., "Ribozymes, riboswitches and beyond: regulation of gene expression without proteins," *Nat Rev Genetics*, vol. 8, no. 10, p. 776, 2007.

[140] Shao, Y. and Bassler, B. L., "Quorum-sensing non-coding small RNAs use unique pairing regions to differentially control mRNA targets," *Mol Microbiol*, vol. 83, no. 3, pp. 599–611, 2012.

[141] Shapiro, B. A., Kasprzak, W., Grunewald, C., and Aman, J., "Graphical exploratory data analysis of RNA secondary structure dynamics predicted by the massively parallel genetic algorithm," *Journal of Molecular Graphics and Modelling*, vol. 25, no. 4, pp. 514–531, 2006.

[142] Shapiro, B. A., Yingling, Y. G., Kasprzak, W., and Bindewald, E., "Bridging the gap in RNA structure prediction," *Curr Opin Struct Biol*, vol. 17, no. 2, pp. 157–165, 2007.

[143] Shapiro, B. A. and Zhang, K., "Comparing multiple RNA secondary structures using tree comparisons," *Computer applications in the biosciences: CABIOS*, vol. 6, no. 4, pp. 309–318, 1990.

[144] Solem, A. C., Halvorsen, M., Ramos, S. B., and Laederach, A., "The potential of the riboSNitch in personalized medicine," *Wiley Interdisciplinary Reviews: RNA*, vol. 6, no. 5, pp. 517–532, 2015.

[145] Sorescu, D. A., Möhl, M., Mann, M., Backofen, R., and Will, S., "CARNA – alignment of RNA structure ensembles," *Nucleic Acids Res*, vol. 40, no. W1, pp. W49–W53, 2012.

[146] Spitale, R. C., Flynn, R. A., Torre, E. A., Kool, E. T., and Chang, H. Y., "RNA structural analysis by evolving SHAPE chemistry," *Wiley Inter Reviews: RNA*, vol. 5, no. 6, pp. 867–881, 2014.

[147] Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., and Giegerich, R., "RNAshapes: an integrated RNA analysis package based on abstract shapes," *Bioinformatics*, vol. 22, no. 4, pp. 500–503, 2006.

[148] Stein, P. and Waterman, M. S., "On some new sequences generalizing the Catalan and Motzkin numbers," *Discrete Mathematics*, vol. 26, no. 3, pp. 261–272, 1979.

[149] Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J., and Doyle, J., "Robustness of cellular functions," *Cell*, vol. 118, no. 6, pp. 675–685, 2004.

[150] Stoltenburg, R., Reinemann, C., and Strehlitz, B., "SELEXa (r)evolutionary method to generate high-affinity nucleic acid ligands," *Biomol Eng*, vol. 24, no. 4, pp. 381–403, 2007.

[151] Sükösd, Z., Swenson, M. S., Kjems, J., and Heitsch, C. E., "Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions," *Nucleic Acids Res*, vol. 41, no. 5, pp. 2807–2816, 2013.

[152] Swenson, M. S., Anderson, J., Ash, A., Gaurav, P., Sükösd, Z., Bader, D. A., Harvey, S. C., and Heitsch, C. E., "GTfold: Enabling parallel RNA secondary structure prediction on multi-core desktops.," *BMC Res Notes*, vol. 5, no. 1, pp. 341–341, 2012.

[153] Tabei, Y., Tsuda, K., Kin, T., and Asai, K., "SCARNA: fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments," *Bioinfor*, vol. 22, no. 14, pp. 1723–1729, 2006.

[154] Tomsey, C. S., Kurtz, M., Flowers, B., Fumea, J., Giles, B., and Kucherer, S., "Case work guidelines and interpretation of short tandem repeat complex mixture analysis," *Croatian medical journal*, vol. 42, no. 3, pp. 276–280, 2001.

[155] Torres, Y., Flores, I., Prieto, V., López-Soto, M., Farfán, M. J., Carracedo, A., and Sanz, P., "DNA mixtures in forensic casework: a 4-year retrospective study," *Forensic science international*, vol. 134, no. 2-3, pp. 180–186, 2003.

[156] Trefethen, L. N. and Bau III, D., *Numerical Linear Algebra*. Philadelphia: Society for Industrial and Applied Mathematics, 1997.

[157] Treiber, D. K. and Williamson, J. R., "Exposing the kinetic traps in RNA folding," *Curr Opin Struct Biol*, vol. 9, no. 3, pp. 339–345, 1999.

[158] Tu, K. C. and Bassler, B. L., "Multiple small RNAs act additively to integrate sensory information and control quorum sensing in Vibrio harveyi," *Genes Dev*, vol. 21, pp. 221–233, 2007.

[159] Tu, K. C. and Bassler, B. L., "Gene dosage compensation calibrates four regulatory RNAs to control Vibrio cholerae quorum sensing," *EMBO*, vol. 28, pp. 429–439, 2009.

[160] Tucker, B. and Breaker, R., "Inventing and improving ribozyme function: rational design versus iterative selection methods," *Curr Opin Struct Biol*, vol. 15, no. 3, pp. 342–348, 2005.

[161] Turner, D. H. and Mathews, D. H., "NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure," *Nucleic Acids Res*, p. gkp892, 2009.

[162] Turner, D. H. and Mathews, D. H., "NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure," *Nucleic Acids Res*, vol. 38, no. suppl 1, pp. D280–D282, 2010.

[163] Turner, D. H., Sugimoto, N., and Freier, S. M., "RNA structure prediction," *Ann Rev Biophys and Biophys Chem*, vol. 17, no. 1, pp. 167–192, 1988.

[164] Voss, B., Giegerich, R., and Rehmsmeier, M., "Complete probabilistic analysis of RNA shapes," *BMC biology*, vol. 4, no. 1, p. 5, 2006.

[165] Wagner, A., "Robustness and evolvability: a paradox resolved," *Proc Natl Acad Sci USA*, vol. 275, no. 1630, pp. 91–100, 2008.

[166] Walter, A. E., Turner, D. H., Kim, J., Lyttle, M. H., Müller, P., Mathews, D. H., and Zuker, M., "Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding," *Proc Natl Acad Sci USA*, vol. 91, no. 20, pp. 9218–9222, 1994.

[167] Weeks, K. M., "Advances in RNA structure analysis by chemical probing," *Curr Opin Struct Biol*, vol. 20, no. 3, pp. 295–304, 2010.

[168] Wilke, C. O., "Selection for fitness versus selection for robustness in RNA secondary structure folding," *Evolution*, vol. 55, no. 12, pp. 2412–2420, 2001.

[169] Williams, A. L. and Tinoco, I., "A dynamic programming algorithm for finding alternative RNA secondary structure," *Nucleic Acids Res*, vol. 14, no. 1, pp. 299–315, 1986.

[170] Wilm, A., Mainz, I., and Steger, G., "An enhanced RNA alignment benchmark for sequence alignment programs," *Algor for Mol Biol*, vol. 1, no. 1, p. 19, 2006.

[171] Wuchty, S., Fontana, W., Hofacker, I. L., Schuster, P., and others, "Complete suboptimal folding of RNA and the stability of secondary structures," *Biopolymers*, vol. 49, no. 2, pp. 145–165, 1999.

[172] Xayaphoummine, A., Bucher, T., and Isambert, H., "Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots," *Nucleic acids research*, vol. 33, no. suppl 2, pp. W605–W610, 2005.

[173] Xu, X., Ji, Y., and Stormo, G. D., "RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment," *Bioinfor*, vol. 23, no. 15, pp. 1883–1891, 2007.

[174] Zhao, X., Koestler, B. J., Waters, C. M., and Hammer, B. K., "Post-transcriptional activation of a diguanylate cyclase by quorum sensing small RNAs promotes biofilm formation in Vibrio cholerae," *Mol Microbiol*, vol. 89, no. 5, pp. 989–1002, 2013.

[175] Zhu, J., Miller, M. B., Vance, R. E., Dziejman, M., Bassler, B. L., and Mekalanos, J. J., "Quorum-sensing regulators control virulence gene expression in Vibrio cholerae," *Proc Natl Acad Sci*, vol. 99, no. 5, pp. 3129–3134, 2002.

[176] Zuber, J., Sun, H., Zhang, X., McFadyen, I., and Mathews, D. H., "A sensitivity analysis of RNA folding nearest neighbor parameters identifies a subset of free energy parameters with the greatest impact on RNA secondary structure prediction," *Nucleic Acids Res*, p. gkx170.

[177] Zuker, M., "RNA folding prediction: The continued need for interaction between biologists and mathematicians," *Lect Math Life Sci*, vol. 17, pp. 87–124, 1986.

[178] Zuker, M., "On finding all suboptimal foldings of an RNA molecule," *Science*, vol. 244, no. 4900, pp. 48–52, 1989.

[179] Zuker, M., "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Res*, vol. 31, no. 13, pp. 3406–3415, 2003.

[180] Zuker, M. and Jacobson, A., "Using reliability information to annotate RNA secondary structures.," *RNA*, vol. 4, no. 6, pp. 669–679, 1998.

[181] Zuker, M. and Jacobson, A. B., "'well-determined regions in RNA secondary structure prediction: analysis of small subunit ribosomal RNA," *Nucleic Acids Res*, vol. 23, no. 14, pp. 2791–2798, 1995.

[182] Zuker, M., Jaeger, J. A., and Turner, D. H., "A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison," *Nucleic Acids Res*, vol. 19, no. 10, pp. 2707–2714, 1991.

[183] Zuker, M., Mathews, D. H., and Turner, D. H., "Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide," in *RNA Biochem Biotechnol*, pp. 11–43, Springer, 1999.

[184] Zuker, M. and Sankoff, D., "RNA secondary structures and their prediction," *Bull Math Biol*, vol. 46, no. 4, pp. 591–621, 1984.

[185] Zuker, M. and Stiegler, P., "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," *Nucleic acids res*, vol. 9, no. 1, pp. 133–148, 1981.

[186] Zuker, M. and Stiegler, P., "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," *Nucleic Acids Res*, vol. 9, no. 1, pp. 133–148, 1981.